Automated Virtual Coach for Surgical Training

by

Anand Malpani

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

February, 2017

Abstract

Surgical educators have recommended individualized coaching for acquisition, retention and improvement of expertise in technical skills. Such one-on-one coaching is limited to institutions that can afford surgical coaches and is certainly not feasible at national and global scales. We hypothesize that automated methods that model intra-operative video, surgeon's hand and instrument motion, and sensor data can provide effective and efficient individualized coaching. With the advent of instrumented operating rooms and training laboratories, access to such large scale intra-operative data has become feasible. Previous methods for automated skill assessment present an overall evaluation at the task/global level to the surgeons without any directed feedback and error analysis. Demonstration, if at all, is present in the form of fixed instructional videos, while deliberate practice is completely absent from automated training platforms. We believe that an effective coach should: demonstrate expert behavior (*how do I do it correctly*), evaluate trainee performance (*how did I do*) at task and segment-level, critique errors and deficits (*where and why was I wrong*), recommend deliberate practice (*what do I do to improve*), and monitor skill progress

ABSTRACT

(*when do I become proficient*).

In this thesis, we present new methods and solutions towards these coaching interventions in different training settings *viz.* virtual reality simulation, bench-top simulation and the operating room. First, we outline a *summarizations*-based approach for surgical phase modeling using various sources of intra-operative procedural data such as – system events (sensors) as well as crowdsourced surgical activity context. We validate a crowdsourced approach to obtain context summarizations of intra-operative surgical activity. Second, we develop a new scoring method to evaluate task segments using rankings derived from pairwise comparisons of performances obtained via crowdsourcing. We show that reliable and valid crowdsourced pairwise comparisons can be obtained across multiple training task settings. Additionally, we present preliminary results comparing inter-rater agreement in relative ratings and absolute ratings for crowdsourced assessments of an endoscopic sinus surgery training task data set. Third, we implement a real-time feedback and teaching framework using virtual reality simulation to present teaching cues and deficit metrics that are targeted at critical learning elements of a task. We compare the effectiveness of this real-time coach to independent self-driven learning on a needle passing task in a pilot randomized controlled trial. Finally, we present an integration of the above components of task progress detection, segment-level evaluation and real-time feedback towards the first end-to-end automated virtual coach for surgical training.

ABSTRACT

Advisor:

Gregory D. Hager, PhD

Mandell Bellmore Professor of Computer Science

Johns Hopkins University

Thesis Committee and Readers:

Russell H. Taylor, PhD

John C. Malone Professor of Computer Science

Johns Hopkins University

C. C. Grace Chen, MD

Associate Professor of Gynecology and Obstetrics

Johns Hopkins University

Carla M. Pugh, MD, PhD, FACS

Susan Behrens, MD Professor of Surgical Education

University of Wisconsin, Madison

S. Swaroop Vedula, MBBS, MPH, PhD

Assistant Research Professor

Malone Center for Engineering in Healthcare

Johns Hopkins University

Acknowledgments

This dissertation in my name has been possible because of the amazing experience I have had at Johns Hopkins University and I owe it to many people – faculty, research, software and administrative staff, collaborators from engineering, surgery and industry, senior, junior and fellow graduate students, undergraduate students, student activity groups, friends, family and my dearest and closest ones.

First, is my advisor and guide, Prof. Gregory D. Hager. He has been a phenomenal person to know and a very supportive advisor throughout my dissertation journey. He encouraged my thoughts about surgical coaching and helped them evolve into scientific contributions through his adept insight about problem solving. He inspired me to always think about the big picture and significance of what I am doing. But, at the same time, he informed me that it can be easy to get lost in complex multi-faceted problems. He advocated breaking down the big picture and ideas into smaller problems which can at times be "low hanging fruits". I thank him for all this and more.

Second, I thank Prof. S. Swaroop Vedula, who came in during my third year and

his contribution is evident from my publications list. We have worked together on several projects and it is his intellectual grasp and understanding of statistics and clinical trials that have helped me perform the various user studies described in this thesis. He introduced me to concepts of statistical tests and heuristics that have shaped me to produce good science. I was fortunate to find a postdoctoral mentor like him to brainstorm on ideas and convert them into scientific contributions to the community along with valid science.

I thank Prof. Russell H. Taylor. It is very important to mention that this thesis was possible due to a project I undertook as part of the Computer Integrated Surgery II course. The simulation sandbox project in collaboration with Intuitive Surgical Inc. gave me a foot in the door to explore virtual reality (VR) simulation. This led to development of the concept of an automated virtual coach using VR training. I had the privilege to discuss my thesis and related projects with Russ during these years and received highly useful feedback that helped develop the various contributions a lot.

I thank Dr. C. C. Grace Chen, who has been the most flexible clinician I have worked with thus far. She has been my clinical advisor on all components of this thesis. She has provided the backing behind all my ventures from VR to bench-top to OR (operating room). It is rare to find such clinical support. She introduced me to the workings of surgical residency programs and patient workflows. Many of the thesis ideas and contributions came to my mind while working with her in project

meetings, training labs and the OR.

Finally, I thank Dr. Carla M. Pugh, who agreed to be a part of my thesis reading committee and provided valuable comments on the thesis work based on her immense experience in surgical education and training. She is also one of those surgeons who embrace such work from engineers like myself and encourage them to keep on going and bring technological advancements to the surgeon's operating table.

I thank the various sources of funding I have received during these years to support my living expenses and tuition fees. I thank the National Science Foundation awards – NRI 1227277, CPS 0931805 and CDI-II 0941362, the National Institutes of Health award 1R21EB009143-01A1 and 5R01DE025265-0, the Intuitive Surgical Inc. Student Fellowship, the Johns Hopkins University Science of Learning Institute research award, the Link Foundation Fellowship for Modeling, Simulation & Training, and the Johns Hopkins University internal funds from departments of Computer Science, Gynecology, Urology and the Malone Center for Engineering in Healthcare.

When I came to Hopkins, I joined an existing project group called "Language of Surgery" and I found myself amazed and intrigued by the problems being addressed. I thank all the project members for their meaningful inputs and guidance in development of my ideas. Special mention to Prof. Sanjeev Khudanpur for questioning assumptions, arguments and conclusions made by me and getting me to think about different aspects of my work by bringing in his vast experience from speech and language processing. Co-researchers play an important role in providing the

scientific ambiance to conduct good research and I thank my colleagues Swaroop Vedula, Narges Ahmidi, Yixin Gao, Colin Lea, Jonathan Jones, Robert DiPietro, Lingling Tao, Piyush Poddar, Nicolas Padoy, Luca Zappella, Benjamín, Béjar, Kelleher Guerin, Amod Jog and Thomas Tantillo for that and much more. Special thanks to Prof. René Vidal and Prof. Rajesh Kumar for providing the initial support and establishing a strong project. In a project like this, success and value is dependent on the existence of a strong foundation of clinical expertise which was provided by Drs. Grace Chen, Masaru Ishii, Jeremy Richmon, Martin Curry, Michael Marohn, Mohamad Allaf, Ardavan Akhavan, Edward Tanner, Amanda Fader, Christian Pavlovich and Jen-Jane Liu. Special thanks to collaborators at the Minimally Invasive Surgical Training and Innovation Center – Drs. Gyusung Lee and Mija Lee. I extend my deepest thanks to a number of staff at the medical school that showed patience and helped me with the multiple data collection studies I was setting up across the hospital – Sue Eller, Josie, Marcy, Lyric, Mary Grace, Siva, Roy Shipley, Darren, Kim Duncan, Institutional Review Board committee members, clinical engineers, network security, facilities, nursing and OR technicians at Hopkins East Baltimore and Bayview campuses.

I have had wonderful lab members at the Computer Interaction and Robotics Laboratory (CIRL), directed by Prof. Hager. I thank Austin, Kel, Haluk, Sharmi, Dan, Rafael, Nicolas, Pezhman, Carol, Narges, Colin, Ayushi, Rob, Amanda, Chris, Xiang, Chi, Jon J, Jon B and people I am definitely missing out on! I thank Amod

ACKNOWLEDGMENTS

Jog, Yixin Gao, Kel Guerin, Tom Tantillo, Qian Zhao and Dan Obenshain for the time spent during the first two years of my arrival at Hopkins and working together at late nights in the Visual Imaging and Surgical Robotics (VISR) lab and Mock OR. Likewise, I have had the great fortune of being part of a larger research group – Laboratory of Computational Sensing and Robotics (LCSR) – which has given me exposure to a huge pool of talented people working on some very cool projects. I thank Profs. Russ Taylor, Peter Kazhanzides, Noah Cowan, Louis Whitcomb, Simon Leonard for this. I thank lab mates Ann, Tom, Xingchi, Sue Sue, Tricia, Manu, Ravi, Subranshu, Wen, Ming, Seth, Francisco, Julian, Ehsan B, Tutkun, Rob G, Rob N, Preetham, Ben F, Javad, Singchun and Molly. I also had an excellent department full of experienced people in different domains of Computer Science (CS). I thoroughly enjoyed the different courses from systems to algorithms to applications. I thank Profs. Misha Kazhdan, Scott Smith, Jerry Prince, Sanjeev Khudanpur, Giuseppe Ateniese, David Yarowsky and Yair Amir. I thank fellow students Hari, Nikita, Tom, Dano, Amod, Puriya, Razieh, Katie, Thijs, Ashis S, Ben S, Farhan D, Rachel S, Benj, Abhinav, Chris W, Jacob. I have learnt a great deal and adopted a lot of knowledge from interactions with these people at CIRL, LCSR, CS and other departments.

Of course, none of the last-minute to-dos would be possible without the flexibility and availability shown by the administrative office to accommodate travel reimbursements, scheduling meetings and events, equipment purchasing, and very importantly, payrolls. I thank LCSR admins – Elisa, Alison, Jamie, Julia, Robert, Lorrie and Rose,

ACKNOWLEDGMENTS

CS admins – Cathy, Debbie, Laura, Tracy, Zack, Jess(ica), Malone Center admins – Nina and Vess. Being an international student away from home, it was the love and care shown by Cathy and Debbie who have been mother-like to me and Alison who has been an elder sister to me, that made this transition easy. I wish them the very best in life.

Coming in from a non-existent computer science background and definitely, a non-existent medical robotics background, it was the great deal of experience from a few people that have transformed me from an almost zero-level programmer and systems engineer to a multi-language developer putting together this coaching framework. I thank Anton Deguet (LCSR), Balazs Vagvolgyi (LCSR) and Simon DiMaio (Intuitive Surgical Inc. – ISI) for all of it. They have provided me endless help in software architecture, framework development, debugging, interfacing with devices, system administration and much more. I cannot thank them enough. I extend my thanks to CS software support staff as well - Steve D, Steve R and Phil.

Almost all my work in the past 7 years has been around the da Vinci Surgical System and my thesis content, experiments and results would not be possible without the ongoing and fruitful collaboration LCSR has had with ISI. I have received a lot of help from collaborators at ISI to work with the da Vinci system for work contained inside and outside of this thesis. I thank Simon, Tony, Henry, Omid, Anusha, Ashwin, Sina, Prasad and Joey. I thank SenseGraphics AB for their code sharing as well as help with the teaching and feedback software development, specially Daniel and Umut.

ACKNOWLEDGMENTS

Hopkins brought to me a pool of resourceful and talented undergraduate student researchers to whom I give a lot of credit for the data curating, processing and labeling. Their work has helped a lot of projects including my thesis at the least. I acknowledge Saranga Arora (Biomedical Engineering - BME) for her contribution to the WarmUp Hysterectomy data set as well as the crowdsourcing summarizations work. I thank Madeleine Waldram (Maddy, Public Health Studies) and Caitlin Romanzcyk (Katy, BME) for their help in cleaning up and annotating of multiple surgical data sets. I thank George Chen (BME) and Anne Evered (Stanford University) for help with annotation of the MultiSite Suturing data set. I thank Srihari Mohan (CS) for setting up the web layout used for the crowdsourcing summarization work. I thank Serena Thomas (BME) for curating the WarmUp study data files.

Life outside of the research laboratory is equally important to find a balance in physical and mental wellbeing. A core group of friends have made the stay at Hopkins and in Baltimore especially comfortable and homely. I thank Siddharth M, Tushar, Mohit, Aagam, Jeet, Anushka, Akanksha, Nandini, Arthi, Isha, Dharini, Sharanya, Srikanth, Smith, Hardik, Omkar, Amod, Pooja, Ramya, Sneha R, Pavan, Vishwa, Geet, Kundu, Princy, Vatsal, Sathya, Smita, Anagha, Aditya, Neha, Ambhi, Rajita, Shambhavi, Sonal, Abdul, Raghu, Varun, Shourya, Abha, Sravya, Gowtham, Ashish, Amit, Vaibhav S, Piyush, Rujuta, Akshay, Vidya, Divya, Amrutha, Farhad, Balaji, Michelle, Samata, Sahil, Arvind, Hari, Madhuri, Kushal, Hansa and so many more names I am forgetting.

ACKNOWLEDGMENTS

I have participated in extracurricular activities during my PhD degree. I thank Kristin Spera and Candice Ford for introducing me to the Johns Hopkins Club Field Hockey team with whom I played a good 4 years. I thank the other field hockey team members – Julie R, Jackie D, Ale S, Laura K, Alex B, Savannah, Rachel Y, Casey M, Tatiana and Megan D. I thank the dance group - Sid, Sarita, Divya, Kanika, Aakruti, Ruchi, Esha and Pranay. I thank the Johns Hopkins Robo Challenge folks – Ann, Tom, Xingchi and Colin. I thank the President's Day of Service staff and volunteers to provide me an opportunity to do community outreach work. I thank the Johns Hopkins Barnstormers student theater group for taking me in as a lighting designer – Raidizon, Kate D, Elizabeth S, Alisson C, Gillian, Megan S, Laura N, Julia, Jon, and Emily S.

Going back to my school, high school and college days, there are many friends, teachers and affiliates who have shaped me and directed me to where I am today. I thank my school friends from Mahindra Academy for staying close at heart and for creating the competing environment that motivated me to keep pushing forward – Satpreet, Sujay, Yasmin and Hitesh. I thank my teachers Mrs. Sawant, Jagannathan, Menon, Vallavde, Geeta, Churi, D'Souza and others for providing me the early education and overall development towards all study disciplines. I thank Tanvi, from my high school for being around all this while and cheering me on as I progressed through college and graduate school. I owe a great deal of my conceptual understanding of physics, chemistry and mathematics to the outstanding tutors whom I was able to

learn from. I thank Profs. Chavale, Amit K, Badve, and Johnny V, Ravi Kishore, Jeeku A, Anshul J and Samir K.

Undergraduate studies at the Indian Institute of Technology Bombay (IITB) opened doors for me to explore a range of activities outside the curriculum and I would like to acknowledge the contribution IITB has had in my personality and character development. It has been far and beyond. I thank the various activity clubs I was able to partake and the students involved in them – Electrified (electronics), TechniC and UMIC (robotics), Rang (fine arts), Fourth Wall (theater), Silver Screen (film and media), InSync (dance), PAF (performing arts festival), Mood Indigo (cultural festival), and the Hostel 8 council. I thank my robotics group "penTium 5" – Mehul, Mukul, Pratik, Smit and Vishnu for the interest I developed in robotics, design, engineering and programming. This group played a huge role to direct me towards robotics research. I thank my batch mates who were part of the group "wing gOOgLe" who brought diversity in culture, talent, hobbies and personalities from around the country and who have continued to stay a well knit group till date – Shagun, Oru, Enthu, Tejas, Shah, Mehul, Smit, Sushant, M, Pandit, Kataria, Singh, Party, Sanket, Babu, Pupun, Birju, Vascool, Vivek, Edla, Reddy, Monjre, Naumzy, Naash and Panji. I am thankful to Siraj and Co., Ashish Goel and Shyam Jade who were the people responsible for reviewing my essays and statements during the graduate school applications. I thank Shruti, GT and other junior students who have shown their pride and faith in me all these years.

ACKNOWLEDGMENTS

I am especially thankful to Amod Jog, Varun Jog and Siddharth M, for they were the ones who introduced me to JHU, LCSR and VISR. This led to my graduate school application at Hopkins. So, none of this would have materialized if it weren't for the chain of emails passed along by these people to me. Thank you! I thank my undergraduate friends who were in the US with whom I explored different cities across the country – Siddharth, Mehul, Pratik, Vishnu, Manas, Mukul, Naumaan, Tejas, Kalyan, Vivardhan, Adith, Adidas and Raunak.

Today I would not be the person I am without the endless efforts and resources that my parents have put in my upbringing and growth. They have given everything and more to let me achieve my education, ambitions and dreams. I thank my mom, Shobha (Mummy) for the science lessons, life lessons and great food, and developing my social and interpersonal skills. I thank my dad, Omprakash (Papa) for the math lessons, finance lessons and great office and cafeteria food, and shaping me into a planner and organizer. I appreciate the belief you both have put in me. Everyone needs a childhood companion and I have been lucky to find that in my brother (Vaibhav). He has played roles of mentee and mentor in my life as we have grown up. He has shared his wisdom and advised me on important decisions as well as helped me bounce off ideas. I thank my brother for always being there. There is no limit on words I can write to thank you all – Mummy, Papa and Vaibhav.

I need to acknowledge Hopkins and the application review committee and a series of fortunate events that led me to apply at Hopkins and a whole bunch of friends who

led me to meet my wife (Princy). I thank Princy's parents for allowing me to marry her. I am belittled by the amount of support and faith Princy has shown in me and towards my work. She has been an ever-helpful friend, fellow researcher, brilliant co-worker, wise counselor, critical proof reader, sound teacher, excellent chef and caring and loving wife. I cannot list the infinitely many roles she has taken in my life. I love you and thanks for being around all along these last few years. I cannot imagine this journey without your presence. I thank Princy's family who have been warm, supportive and welcoming. Thank you – Radhika, Juhi, Chaitanya, Ramkrishna, Mummy (Bharti) and Papa (Navin).

While I have taken the space and time to list all the above acknowledgments, I have definitely missed out on very important names and I apologize for that. If only I knew I would have kept a record of all the people I met. I would like to end this section by saying that words cannot describe how I feel right now and how I felt on the day of my dissertation defense. I feel humbled and fortunate that I have had so many people contribute towards this thesis in different ways. I have thoroughly enjoyed this journey and I am quoting Henry Lin (PhD, CS, Johns Hopkins University, 2010) here: "I am a happy man."

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Thesis Statement

Multi-modal data obtained from surgical interventions can be modeled using crowd-sourced representations and machine learning techniques to deliver automated virtual coaching capable of providing relevant, targeted, critical and individualized learning in a virtual reality environment.

While major contributions of this thesis are applicable to many areas wherein automated training and coaching are needed, we talk about our developments and inventions in the context of surgical education.

1.2 Motivation

1.2.1 Surgical Education: lack of regular testing

William S. Halstead proposed the first residency program in United States at the Johns Hopkins University with a model of "see one, do one, teach one" in 1890. Such learning on patients in the operating room (OR) conflicted with the notion of providing them the best care possible. This was excusable in that era of surgical education when simulation technology was lacking. However, it was only at the dawn of 21st century that surgical educators and policy makers targeted this model of learning, and deemed it unfit and not in line with the "patient first" model of health care providers. In 2003,[3] the Accreditation Council for Graduate Medical Education (ACGME) moved surgical training to a competency-based lifelong learning model from Halstead's time-based development of career lasting skills. This has led to the inception of competency-based testing and certification programs like Fundamentals of Laparoscopic Surgery (FLS),[4] Fundamentals of Endoscopic Surgery (FES),[5] and Fundamentals of Robotic Surgery (FRS).[6] As per guidelines laid down by the different surgical societies, surgeons are required to get re-certified only every 10 years. But, what about their case-by-case feedback, day-by-day assessment, week-by-week learning or month-by-month competency?

Current surgical training has no room and mechanism for such regular testing of skills development and maintenance.

1.2.2 Coaching: why not in surgery?

Surgery is a performance. Athletes and musicians must perform and give their best – physically and mentally – at each competition and concert. Quite similarly, a surgeon must put up their best performance for each patient in the operating theater. However, there is a striking difference between the regular routine of surgeons compared to athletes and musicians. Most of sporting and music professionals continue to have their individual or team coaches even after showing immense talent and capacity to perform above and beyond the norm. The notion of coaching in surgery, on the contrary, is quite alien. In fact, having a coach may be perceived as signs of being incompetent and may lead to loss of authority as subjects reported in the study by Mutabdzic et al.[7] Even at highest level of competition, the Olympic Games, a badminton player looks up to their coach and has a chat in between games, for feedback and tactics. Their coach has conditioned them and become their outside eyes and ears to tell them where they are falling short.[8]

A coach can take the form of a teacher and mentor for amateurs (residents) and the role of an observer and critique for experts (attending surgeons). Coaching can provide individualized learning and targeted feedback, that traditional standardized-test based training cannot. Coaching focuses on finer aspects of performance and leads to piece-by-piece improvement of skills. But, a coach has to put in considerable amount of time and effort to achieve this. This is possible for professional sports and music coaches who do this as a full-time job, but can be a large limiting factor for the

feasibility of surgical coaching. Outside of operating schedule, clinic duties, research and administrative work, surgeons cannot find time to teach residents as well as coach peers at an individual level. Funding for such coaching-based learning can also be considered an unnecessary financial burden at most medical institutions. Finally, surgical coaching can be subjective and biased to conform to the coach's preferred operating styles, room setup, and similar factors. This can lead to a sense of loss of control to the surgeon being coached.[7] Perceived incompetency, lack of coaching time, scarce finances, loss of autonomy and subjective nature are all deterrent to the concept of manual surgical coaching.

Manual one-on-one coaching has shown effectiveness[9] in improvement of technical skills and reduction in number of errors committed, however its scalability is limited by the current culture and resources in surgery.

1.2.3 Computer Integrated Surgery: easy data access, complex to process?

Computer integrated surgical (CIS) systems have gained wide acceptance in the surgical community to provide assistive technology to surgeons, nurses and other OR staff. Development of CIS devices has led to instrumentation of the OR. Collecting intra-operative data from sensors on these devices, that are increasingly getting integrated in the OR, has become feasible without disrupting workflow. Surgical data in

the form of intra-operative video, instrument and surgeon's hand motion, and sensor signals like cautery usage and irrigation bag weight can be captured courtesy of the installed CIS systems. Gathering large scale data spanning across surgeons, surgical specialties, hospitals, states and countries is becoming plausible.

Alongside, advancements are being made in the domain of machine learning and artificial intelligence to process data sets of large size and variability. Researchers in natural language, speech, image and video processing as well as motion planning and robotics have shown promising results in analyzing and extracting information from text, audio, video and motion data. Over the last decade, computer scientists and engineers have adapted techniques from these areas onto the surgical domain. Representation learning and signal processing methods can transform surgical data into features useful for skill assessment and activity recognition. Computer vision algorithms are able to detect surgical objects and parse the surgical scene to extract contextual cues.[10,11] Pattern recognition and matching techniques are able to look-up similar performances from other surgeons to provide useful guidance and teaching.[12,13] Concepts like virtual fixtures and haptic feedback are able to guide surgeons' motions to confine them to safe zones or push them away from crucial structures like a major artery.[14,15] Human machine collaborative systems for shared automation of repetitive and trivial tasks are being demonstrated as well.[16]

While such computer-assisted intervention (CAI) methods and pipelines are being developed, there are still limitations and shortcomings. None of these technologies

have shown success at a broader scale in terms of multiple surgical specialties or multiple hospitals. Solutions for fine-grain activity recognition in in-vivo data have recently appeared on the horizon and will take time to become as mature as their counterparts in speech and language communities. Automated skill assessment is quite a long way from becoming useful; with current approaches providing skill classification (whether the surgeon is novice v/s intermediate v/s expert) as compared to skill evaluation (assigning a score or standing to a surgeon). Other components of training like feedback and demonstration are completely missing.

While surgical data access in the OR and training labs has become relatively simpler and less disruptive, surgical data science methods are still in their nascent stage to deliver surgical activity recognition, surgical skill assessment, teaching with feedback and demonstration.

1.2.4 Crowdsourcing: harnessing the wisdom of crowds

Surgical data analytic tasks, which are still complex for computers to perform, may not be as challenging for a group of humans with basic intelligence. Given a group of humans with average intelligence and structured tasks to perform, one would expect a reliable collective outcome from such a committee.

Crowdsourcing, etymologically "outsourcing" of tasks to a "crowd", has gained

popularity in the last decade or so, for providing solutions using collective intelligence and wisdom of individuals without experience of the problem at hand. The ability of humans to solve problems, and at a scale of more than 3.5 billion internet users around the globe has opened doors to a farm of computing, creative, decision-making and economic resources, at levels never imagined before. The success for crowdsourcing lies in a mutual benefit – the crowd motivated through monetary, social, and/or learning benefits, while the crowdsourcer harnesses intelligence, imagination and investment from a large and diverse pool of humans. Crowdsourcing has been applied in various tasks *viz.* image categorization, fashion critique, product ratings, creating designs, funding projects, digitizing paper documents, cataloging large inventories, security, monitoring, search for survivors in natural disasters, and has impacted many industries like transportation, hospitality, food, employment, design, lending, venture capital, non-profit organizations like Federal Bureau of Investigation, academic research, to name a few.

Recently, biomedical applications have benefited from crowdsourcing complex tasks which current computers are not able to perform and at speeds that surpass experts in the field. The web-based game 'FoldIt' has led to the discovery of native protein structures by the efforts of more than 57,000 internet players.[17] Similarly, the DREAM challenges (Dialogue for Reverse Engineering Assessments and Methods) have been successful in crowdsourcing solutions for Amyotrophic Lateral Sclerosis (ALS) progression models[18] as well as prostate cancer patient survival prediction mod-

els.[19] Likewise, healthcare applications of solving complex medical cases (CrowdMed), providing pathological data analysis,[20] diagnosing malaria infected blood cells,[21] helping substance abuse patients, and analyzing healthiness of food[22] have demonstrated success with crowdsourcing. Overall, there has been a widespread adoption of crowdsourcing across multiple disciplines to 1) solve problems, 2) vote, rate and label data sets, 3) create solutions and designs, and 4) fund projects in need of financial support.

Crowdsourcing has started appearing in surgical data analysis as well. The wisdom of crowds has shown high correlation and agreement with expert skill assessment of surgical training tasks and OR procedures.[23,24] Crowds have generated comparable evaluations of aesthetics of post-plastic surgery patient appearances.[25] Medical image annotation for polyp detection in endoscopic data,[26] abnormality in retinal microscopic fundus data,[27] laparoscopic instruments segmentation in minimally invasive surgical data[28] are some other success stories of crowdsourcing surgical data.

Crowdsourcing has shown potential to generate large scale training data corpuses for the development and validation of CAI technologies with applications in automated and objective surgical skill and activity analysis.

1.2.5 Simulation-based Training: taking learning outside the OR

With the technology revolution we are witnessing in current times, another area relevant to surgical training that has seen advancements is virtual reality (VR) technology. Computer graphics and computing power along with concepts of game design and physics simulation have improved significantly over the past decade. Educators have propagated use of VR in surgical training since 1993.[29] Only in 2002, the first study[30] was conducted to show effective transfer of skills developed from VR simulation-based training to OR; with positive outcome of decreased operating time and lower number of errors. Around the same time, in 2003,[3] ACGME cut down the work hours for residents to 80 per week from a 110 per week[31] so that patient safety would improve with less fatigued residents attending to them. Incidentally, this had an adverse effect and reduced resident's OR experience.[32] Furthermore, OR time has become expensive, health care payer's focus on medical errors has increased, and this combined with the ethical issue of learning on the patient while delivering the best health care possible, have all led to an increased need for simulation-based training.

Simulation, whether physical or VR (computer-based), presents surgical education with a lot of advantages over traditional training in the OR. Firstly, standardized and structured training curricula can be developed using simulation, to train and certify residents and surgeons around the world. A trainee's performance and skill progres-

sion can be tracked, and increasingly complex and difficult training tasks can be introduced based on this learning progression. Secondly, simulation may be customized to expose trainees to extreme scenarios as well as new operating styles without any risk to patients. Errors in simulation are excusable and may result in better surgical proficiency[33] compared to traditional training where patients may get harmed. For example, incorporating errors in training had positive effect on skill retention in central venous catheter placement.[34] Thirdly, simulation labs offer a lot more flexibility for scheduling, and trainees can fit it in their schedule without compromising on OR and other clinical experience. Trainees can acquire basic skills in simulation labs and use attending surgeon's time in OR effectively for skills associated to real patient care and experience. Thereby, also, resulting in more efficient OR workflow and more revenue for the hospital. Finally, simulation-based training transfers well to OR environments as shown in studies conducted to test the hypothesis. Simulation-trained residents had a significantly lower amount of central line infections.[35,36] Similar outcomes have been demonstrated in OR procedures (hernia repair)[37] and emergency care (neonatal)[38,39] as well. In summary, the value of simulation in surgical training is great and has been proven in some cases.

Simulation brings standardization, customization, OR efficiency, and better outcomes to the table compared to traditional or no training[40].

1.2.6 Virtual Reality Training: closer to automation, not there yet!

VR simulation adds on to the list of pros discussed above by providing reduced costs, around the clock availability, and possible automation. Cost is an important outcome of measuring effectiveness of simulation-based training. As mentioned previously,[40,41] current studies did not include a complete cost-effectiveness outcome. Nevertheless, VR training fares well compared to bench-top simulation over multiple cost categories of "equipment and materials" (training materials, durability of materials), "personnel cost" (staff fee, staff/faculty time, administrative staff fee), "facility costs", and "client inputs". VR simulators do not require any consumables or staff to setup the training modules, which takes away administrative costs associated with purchasing and payroll as well. VR simulators do not necessarily require faculty/instructor's time which also leads to lower loss of clinical revenue due to absence of staff from work for teaching. In addition to cost, availability of VR simulators can be 24/7 as it does not have any staffing requirements. VR simulators allow lifelong learning in a private and autonomous fashion which can be a big plus for the surgical coaching agenda.[7] Using internally available task execution information, the computer generates performance metric scores (automated), which are based on previously defined and validated formula and code (objective), and on the same training task parameters and setup as for any other trainee using the same simulator

(structured).

Thus, VR simulation enables automated, objective and structured skill assessment, in the true sense of the phrase.

While VR surgical simulators present such benefits for skill acquisition, voluntary training using them has seen poor response.[42–44] Prior to duty hour restrictions, residents gave 'lack of time' as a reason. In a more recent study, incentivizing simulation-based training with hands-on operating experience as a reward showed poor response (44%) as well, with participants giving personal reasons for lack of frequent practice. Overall, the residents recommended that mandating such training, requiring certain level of proficiency for OR experience, and regular reviewing of performance may improve usage of simulation-based training. An important feedback from the trainees was that simulator training was not similar to exposure from OR experience.

We think that OR experience has two components: real patient experience and faculty surgeon mentoring experience. VR technology has made way to full procedure simulations including flexible organ simulation and soon the face validity of procedure simulations will be satisfactory. This will address the real patient experience component of OR learning. **However, none of the available VR simulators provide any sort of mentoring/coaching that the OR environment provides.** For example:

- Automated and objective, but only "global" assessments are available. There is no breakdown of performance ***evaluation*** at a sub-task level.

- Superficial textual feedback is presented, but just to prompt the trainee about task protocol and steps. Constructive and immediate ***feedback*** on errors and skill deficits is not available.
- Trainees are shown instruction videos, but again, to show a fixed example of task protocol execution and not for demonstrating expert skills. Context-relevant ***demonstration*** using an ideal performance from an experienced surgeon is missing.
- Skill development is tracked by current VR simulators, just to show learning trends, without any suggestions for improvement or indication of graduation.
- Individualized ***deliberate practice*** and ***skill progression*** focused on performance trend are not available on any simulator yet.

All the above missing elements – fine-grain performance evaluation, constructive and immediate feedback on errors and deficits, context-relevant demonstrations, individualized deliberate practice sessions along with monitoring of trainee skill progression – are characteristics of a successful and effective coach.[8,45–48] **Current VR simulators lacks effective coaching.**

1.3 Coaching Activities

We believe that a coach should perform five core activities (Figure 1.1) and provide the corresponding interventions to answer the trainee's questions:

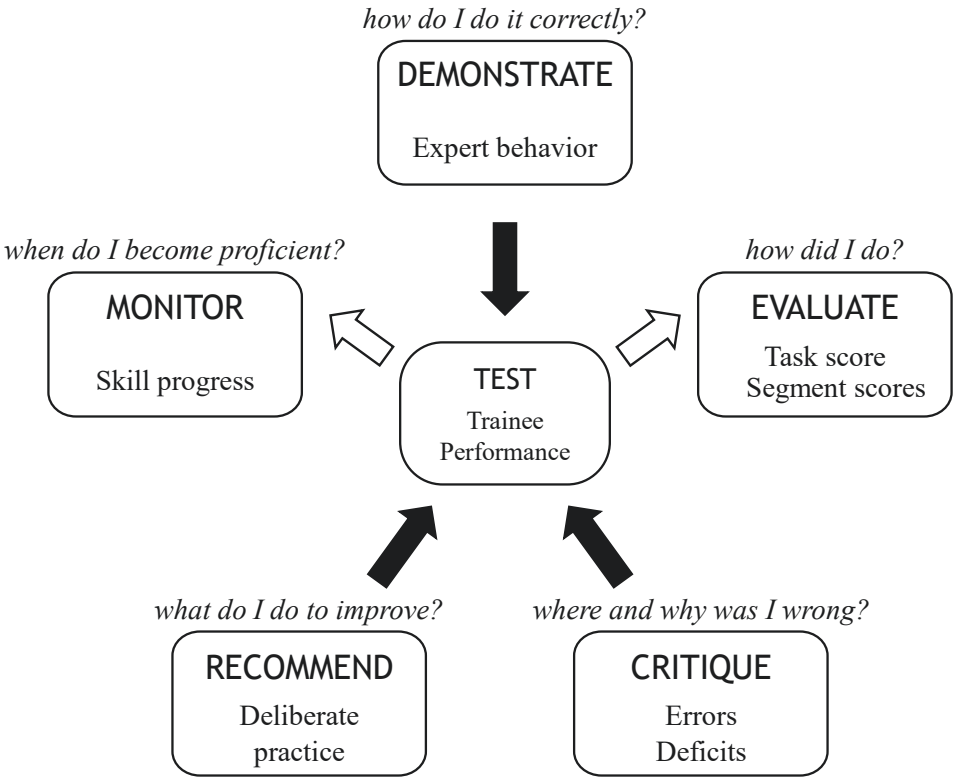


Figure 1.1: Five core coaching activities and corresponding interventions directed to respective trainee questions. The arrows represent information flow from trainee to coach and vice versa.

i) *how do I do it correctly* – `DEMONSTRATE`

ii) *how did I do* – `EVALUATE`

iii) *where and why was I wrong* – `CRITIQUE`

iv) *what do I do to improve* – `RECOMMEND`

v) *when do I become proficient* – `MONITOR`

1.3.1 Current State of Coaching Activities

As discussed before, the current state of surgical training and education is limited. Different research studies have shown the promise and success of such coaching interventions at discipline-wide or institution-wide settings.

`DEMONSTRATE`: The process of illustrating expert behavior and performance is restricted in current training platforms. A fixed set of illustrations and video recordings are used in them with the goal of providing instructions rather than teaching skills. Demonstration of errors in performance is non-existent with few studies[34,49] showing value for error training. The dual console feature of the *da Vinci*® Surgical System enables real-time demonstration wherein the teacher takes over the control of robot from the trainee to show how to perform the current step correctly. Very recently, studies[50,51] have explored the notion of ghost tool overlays for proctoring. However, no quantitative results on their effectiveness in skill development have been studied yet. Recent works[12,13] on video retrieval have shown success in fetching context-relevant videos with the goal of coaching, but haven't looked at impact of learning. Finally, haptics-based virtual fixtures for surgical assistance[15] is an active area of research but so far, no work has used them for teaching surgical skills. In summary, current solutions do not provide context relevant demonstrations to trainees based on their skill level.

`EVALUATE`: A plethora of prior research has focused on the question of surgical skill evaluation. In brief, the current standards are check-list based tools[1,2,52] for assigning

global rating scores (GRS). A lot of variants specific to procedures and disciplines have been developed and validated as well. In addition to these manual approaches that are limited by surgeon (rater) time, automation using quantitative assessments on surgeon hand and instrument motion data[53–55] have been validated. Machine learning and data modeling techniques like hidden Markov models,[56–58] support vector machines[59] as well as string-motif based,[60,61] wavelets-based[62] and video-based[63,64] representations have been explored and tested for surgical skill categorization. All of the above works have focused on global (skill assessments) assigning skill labels[60] or skill scores[65] at the task-level with no evaluation of segment-level performances. Thus, an instructor or teacher is needed to give verbal evaluation or a breakdown of the performance. In summary, current skill assessment (manual and automated) is limited to global rating scores and do not provide segment-level performance evaluation.

`CRITIQUE`: There have been a fair share of studies to investigate the effectiveness of providing feedback to trainees on their performance. Most of these works[66–71] tested a similar setting wherein an external surgeon (expert) gave feedback post-completion of performance. Additionally, concurrent feedback by experts[72] and peer feedback[73] have shown success at small scales. Like skill assessment, such manual feedback requires the critic's time. Automated feedback methods have been absent so far in the literature. Some feedback mechanisms like visual force feedback[74] and virtual fixtures-based guidance[15] have demonstrated improvement in task execution but do not address the notion of explaining errors and deficits in skill. In summary, critiquing

of performance to point out reasons and compliment the performance evaluation is still absent in current training platforms.

`RECOMMEND`: Focused practice has been shown to improve skill in domains like performance sports and arts. The concept of deliberate practice has shown initial success in previous studies[46,75] in the domain of surgical education. The activity of recommending deliberate practice relies on the ability to evaluate and critique performance at task and segment-level which is in the nascent stage of development in automated training solutions. In summary, automated recommendation of deliberate practice is not present and manual deliberate practice has been tested by one research group only.

`MONITOR`: In cases where a dedicated training lab instructor is available, a trainee's performance is tracked and corresponding feedback and recommendations are made. Progression through the different levels of proficiency and difficulty are based on the subjectivity and bias of the instructor. The other option wherein such a full-time instructor isn't available is a self-learning setting using VR simulators. But, it has been shown that surgeons are not good at self-assessments.[76] In summary, objective and automated monitoring of trainee progress isn't plausible given current state of the other coaching activities.

It is evident that the availability of a coach performing these core activities is limited and present only in certain research study settings.[9] Different works[68,69,77,78] have indicated the need for an automated solution towards these coaching interven-

tions. Manual coaching, if available, faces limitations as mentioned before.[7] To deliver the above specified coaching activities, an automated coach should be able to perform:

- ✔ scene parsing and context extraction,
- ✔ activity analysis and recognition,
- ✔ assessment of skill at segment and task level,
- ✔ detection of errors and skill deficits,
- ✔ retrieval of relevant demonstrations from a library, and
- ✔ identification of plateauing and peaking.

The current state of research in these domains is highlighted in Table 1.1 along with contributions presented in the remainder of this thesis. We will go into more detailed background and review of existing works in each of these topics in respective chapters.

Table 1.1: State of automated methods required for various coaching activities in the different training environments.

Training Type	Activity Recognition	Skill Assessment	Feedback	Demonstration
VR	✔	🎓[4,5]	🎓[5]	🎓[5]
Bench Top	✔	🎓[4]	✖	✔
OR	✔ 🎓[3]	✔	✖	✖

✔: indicates previous work in literature exists, 🎓[X]: indicates a solution is presented in Chapter "X" of this thesis, ✖: indicates no prior work exists and none is presented in this thesis.

1.4 Contributions

In this thesis, we present novel solutions to the problems of (1) surgical activity modeling, (2) surgical skill assessment, and (3) real-time feedback and teaching, towards the end goal of automated coaching for surgical training. Our specific contributions for each problem are listed below.

Surgical Activity Modeling

We show that operating room procedures can be jointly segmented and classified into constituent surgical steps using summarizations based on activity context, tool and surgeon's hand motions, and system events information.

Surgical Skill Assessment

We propose a new segment-level skill assessment score using a ranking approach based on pairwise comparisons of matched performances.

Real-time Feedback and Teaching

We demonstrate the feasibility of a real-time feedback framework that presents teaching cues and validate its effectiveness in skill learning in a randomized controlled trial.

Automated Virtual Coach

We present an end-to-end automated coaching framework for virtual reality training in surgery that demonstrates, evaluates, critiques, recommends and monitors a trainee's performance and practice.

1.5 Organization

We have partitioned this thesis into chapters corresponding to each of the contributions listed above.

Chapter 2: Data Sets and Terminology

A large amount of effort during this thesis was devoted to collect data sets used in validation experiments that support the thesis's contributions. First, we will introduce the surgical devices and software that enabled this data collection. Then, we will present a summary of surgical data sets obtained from performances during VR training, bench-top simulation training, cadaver lab training and operating room procedures on patients. Detailed data set descriptions will be presented in the appendix. Later, we will define the common terminology and concepts belonging to crowdsourcing and motion data analysis that will be used in this thesis.

Chapter 3: Surgical Activity Modeling

We will begin by introducing a decomposition of surgical activity based on the use of semantic vocabularies. We will describe the different surgical data modalities available and compare the information contained in them in relevance to activity recognition. We will review current techniques for activity modeling by the surgical scenario - bench-top simulation and OR procedures. We will present a pipeline for surgical phase recognition in OR based on summarizations. We will show that summarizations using system events information contain surgical phase information. We will prove that summarizations of surgical activity and context can be obtained with high reliably and validity from a crowd of surgically untrained individuals. We will present a pilot validation of surgical phase recognition using such crowdsourced context summarizations.

Chapter 4: Surgical Skill Assessment

We will review current literature and methods for skill assessment and motivate the need for segment-level evaluation of skill. We will outline our pipeline to produce segment- and task-level skill scores using pairwise comparisons of performances. We will establish reliability and validity in pairwise comparisons of surgical task segment performance videos using the collective wisdom of a surgically untrained crowd. We will prove that relative ratings from crowds are more consistent and reliable compared to absolute ratings for segment-level skill assessments. We will show equivalence

between scores obtained from crowdsourcing- and expertsourcing-trained machine learning methods. We will validate our framework across multiple training paradigms and data sets.

Chapter 5: Feedback and Teaching

We will review previous studies testing the effectiveness of feedback on learning and motivate the need for error analysis in training. We will present new performance analysis metrics based on skill deficits and errors. We will present teaching cues directed at important learning elements of a skill. We will show feasibility of calculating the deficit metrics and presenting teaching cues in the context of needle passing skill in VR simulation. We will demonstrate the effectiveness of real-time feedback using teaching cues in VR training in a randomized controlled trial setting.

Chapter 6: Towards an Automated Virtual Coach

We will introduce the concept of deliberate practice and present its application in surgical proficiency development. We will describe recent validation studies on the effectiveness of coaching and motivate the need for automated surgical coaching. We will list the core activities and corresponding interventions of our proposed automated virtual coach (VC). We will outline the flow of a trainee through the coaching cycle and introduce different components of our VC framework. We will present a detailed process diagram for each of the components in detail. We will layout the first end-

to-end automated coaching system for surgical training and introduce the software implementation so far.

Chapter 7: Discussion and Conclusion

We will list the contributions of this thesis in the domains of crowdsourcing, surgical activity modeling, surgical skill assessment, error- and deficit-based feedback and surgical coaching. We will present a summary for each of the chapters and the contributions made in them. We will list current limitations of the work, future questions to be answered, and new areas of research related to surgical training and coaching. We will talk about future development, experiments and validations of the automated virtual coach (VC). We will discuss the philosophy of surgical data science and its role in shaping the future of surgical training, specifically, and surgical practice, in general.

Chapter 2

Data Sets and Terminology

We have set out to develop techniques that can deliver the different activities of an automated surgical coach viz. demonstrate expert behavior, evaluate segment- and task-level performance, critique errors and deficits in performance, recommend deliberate practice and monitor skill progress. A common requirement for these coaching actions is a pool of data that is large, broad in task (patient) complexity, representative of the surgeon population, inclusive of all skill levels, and consistent and accurate for retrieval and storage. Modeling techniques from machine learning rely on the availability of rich training data sets to learn the representation, recognize the patterns and predict the outcomes/properties of new data at hand. This has led to development of tools for data acquisition and pre-processing, and curation of large data sets in the communities of speech recognition,[79,80] natural language processing,[81,82] computer vision,[83,84] medical imaging[85] (Alzheimer's Disease Neuroimaging

Initiative (ADNI) database) and so on.

Data mining and big data analytics in surgery (more generally in healthcare) are gaining attention. Unlike the computer vision and speech communities, healthcare data procurement faces a lot of challenges. For example, missing data in patient healthcare records and outcome measures is common.[86] Another challenge is the associated privacy and legal concern with obtaining healthcare data. Hospital administration are concerned about the liability associated with intra-operative data capture due to the potential of malpractice claims over post-operative complications. In addition, a data collection protocol needs to go through the ethics board review for human subjects research. Any recording equipment that is added to existing infrastructure requires inspection from clinical engineering personnel. In the OR, data recording equipment should be a non-disruptive addition to the already complex workflow and requires approval from the OR staff. In summary, the data collection must be invisible in the array of things present in the environment. To add, solving this process at one location (hospital) does not guarantee a success at any other location (even) within the same healthcare system. Several factors including, but not limited to, spatial layouts, approvals, protocols, policies and equipment can change between locations. This makes collecting intra-operative data from multiple sites even more difficult and challenging.

The need for large data sets with set standards is rising along with the need for automated surgical data modeling techniques. A variety of applications including

surgical training and performance feedback require automated methods to scale up, and be effective and efficient. Concurrently, miniaturization and precision of sensors have crossed the bounds of imagination from a decade ago. Computer Integrated Surgery (CIS) devices embedded with such sensors are making way into and getting integrated within the OR infrastructure. Future ORs are being conceptualized with data capture, analysis and presentation at center of the design.[87,88] The feasibility of non-disruptive and scalable intra-operative data recording is becoming easier and within reach day-by-day.

The domain of 'surgical data science'[89,90] is in a nascent stage with open access, standardized large data sets as an important, required and emerging subdomain.

In the first half of this chapter, we will introduce the surgical devices, describe a data collection framework, and summarize different surgical data sets that were collected and curated to validate the contributions presented in this thesis. Later, we will focus our attention on the terminology and concepts that will be used across the different coaching components presented in the chapters that follow.

2.1 Surgical Devices

Surgery is an ancient science wherein the operator (surgeon) manipulates patient anatomy with the goal to inspect, repair and/or remove unwanted growths or infected

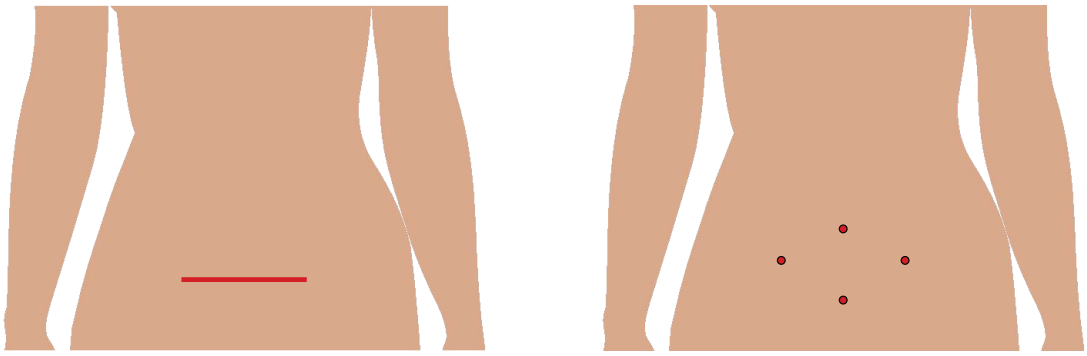


Figure 2.1: Incision sizes in open and minimally invasive surgery (MIS). Red line shows a cartoon incision in case of open surgery like laparotomy. Red circles show the multiple small incisions in case of MIS like laparoscopy.

(diseased) organs. Traditionally, surgery was performed in an "open" fashion by making a large incision on the body to gain access to the internal cavity and organs. For example, laparotomy involves opening the body through the belly to operate in the abdominal cavity. Typically, open surgery leaves the patient with a large post-operative recovery period and requires special care to avoid chances of infections.

Minimally Invasive Surgery

In the late 1980's,[91] the term 'keyhole surgery' came about to be used for describing the new and exciting domain of less invasive or minimally invasive surgery (MIS). MIS typically involves creating a few and small incisions to get access to the anatomy of interest via some form of image guidance (Figure 2.1. For example, laparoscopy is performed via two or more incisions on the abdominal wall to operate in the abdominal cavity in a MIS fashion. Typically, a surgeon holds two thin (chopstick-like)

instruments that enter through two separate incisions, while an assistant manipulates an endoscope inserted through a third hole to provide visual guidance projected onto a monitor in front of the operating surgeon. Immediate advantages of the MIS approach over open surgery are reduction of recovery (rehabilitation) time, less post-operative trauma, better cosmetics and lower chances of infections due to smaller incisions. MIS was adopted by multiple surgical disciplines for various suited procedures – prostatectomy (urology), hysterectomy (gynecology), aortic aneurysm repair (cardiac), to name a few.

The value of MIS was recognized immediately and was high for all the patient benefits resulting from it. However, it had major drawbacks for the surgeon compared to the open surgery. These were multi-fold – loss of direct visualization, no 3D and depth perception, lack of force sensation, and loss of dexterity using human wrist and fingers. Additionally, due to fulcrum effect, controlling the chop-stick instruments was inverted. One had to move the holding end of the instrument left for the operating end to move right. All of this led to a steep learning curve for MIS.

Robot-assisted MIS

Concepts in robotics were starting to develop and could provide solutions for these shortcomings in MIS. In the early 1990s, advances were being made in medical and surgical robotics, with the ROBODOC system (orthopaedic surgery),[92] the PUMA robot (neurosurgery)[93] and the NeuroMate (neurosurgery) showing early suc-

cess and benefits. The Automated Endoscopic System for Optimal Positioning (AESOP) robot[94] was the first Food and Drug Administration (FDA) approved robot in MIS procedures (Computer Motion Inc.). It reduced tremors and improved camera navigation as an assistant to the surgeon. Owing to a funding opportunity by the Department of Defense (DoD) and North American Space Association (NASA), tele-robotic systems were developed to perform precision tasks, specifically surgery in battle fields and in outer space. This led to the development of the da Vinci® Surgical System (dVSS; Intuitive Surgical Inc., Sunnyvale, California) and the Zeus™ system (Computer Motion Inc.), which was later merged with Intuitive Surgical Inc. in 2003. This form of surgery enabled by a tele-robotic system is referred to as 'robot assisted minimally invasive surgery' (RAMIS). We will explain the different components of the dVSS below.

2.1.1 da Vinci® Surgical System

The dVSS was introduced in 1999 as a tele-robotic surgical system to perform MIS. Since 1999, four generations of dVSS systems have been introduced commercially. Here, we will describe the "Si" model which is the third generation of dVSS. A setup of the dVSS Si is shown in Figure 2.2. The operating surgeon sits at the surgeon side console (SSC) while controlling the instruments mounted on the patient side cart (PSC) and viewing the stereo video feed sent through the vision cart of the patient side workspace.

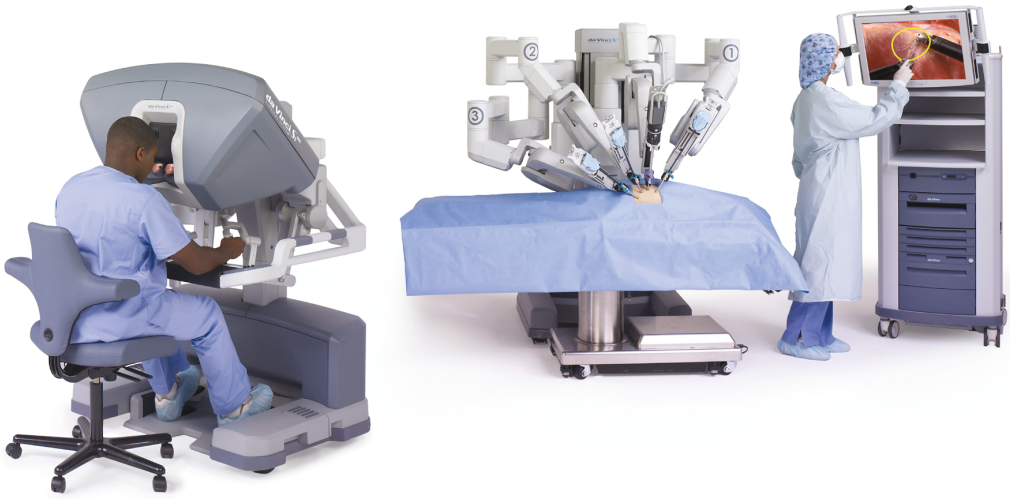


Figure 2.2: dVSS Si version (left to right): surgeon side console (SSC), patient side cart (PSC) and vision cart. The operating surgeon sits at the SSC and controls the instruments mounted on the PSC arms using the joystick controllers on the SSC. The surgeon can view the surgical workspace at the PSC through the stereo viewer that is fed a live, high definition view from the stereo endoscope mounted on PSC via the vision cart. *Image: © 2017, Intuitive Surgical Inc.*

Surgeon Side Console (SSC)

SSC consists of a stereo viewer, master tele manipulators (MTMs) and a foot pedal tray as shown in Figure 2.3. The stereo viewer displays a live, high definition and 3D video feed of the surgical workspace from the PSC endoscope providing the lost depth perception from traditional MIS back to the surgeon. Two MTMs are provided as joysticks for the surgeon to control the robotic arms on the PSC. MTMs are serial robotic arms with eight joints wherein the surgeon controls the end-effector of the instrument using their thumb and index/middle finger (Figure 2.4). Each of the MTMs is also equipped with a pull-back type button referred to as "clutch", that can be triggered using the free index finger on either hand. The clutch disengages the master (SSC) from the slave (PSC) to enable re-positioning of the MTMs when the SSC workspace limits are reached. A foot tray containing a few useful pedals is present. The two black pedals on the left side are for "clutch" (top) and "camera control" (bottom), while the two set of yellow and blue pedals activate the monopolar and bipolar energy devices hooked up to the system via the vision cart (yellow for cut energy type and blue for coagulation energy type). Another black pedal (not visible in Figure 2.3) on the left side wall of the tray is present to allow swapping control to the third instrument arm.

Figure 2.3: Surgeon side console of the dVSS Si version. *Image: © 2017, Intuitive Surgical Inc.*

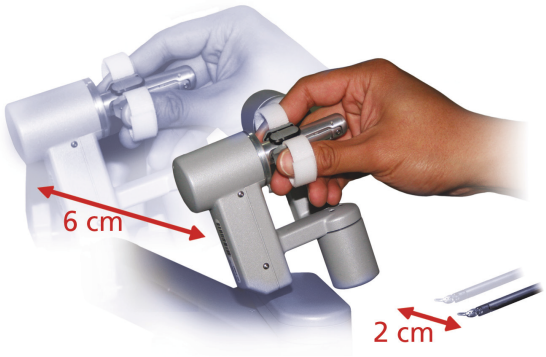


Figure 2.4: Master Tele Manipulator (MTM) of the dVSS Si version *Image: © 2017, Intuitive Surgical Inc.*

Patient Side Cart (PSC)

Four robotic arms - each consisting of a serial link manipulator referred to as the patient side manipulators (PSMs) and the endoscopic camera manipulator (ECM) - are present on the PSC (Figure 2.5). Three of the four robotic arms (the ones numbered) are the PSMs and can hold an instrument, while the center arm is the ECM and holds the high definition stereo endoscope. PSMs and ECM contain four passive joints referred to as *setup joints* that can be adjusted by pressing "setup clutch buttons" and moving the arms by holding them physically. Following this, there are seven active joints on the PSM and four active joints on the ECM that can be controlled using the MTMs on the SSC (Figure 2.6). PSMs and ECM have a trocar holder and instruments and endoscope are passed through a trocar to enter the patient's body. Each PSM and ECM features a *remote center of motion* (RCM) that is defined as the point in 3D space around which the entire motion of the instrument occurs. Basically, the roll, pitch and yaw axis of the PSM and ECM are concurrent at this 3D point. The RCM provides an additional benefit over traditional MIS. Placing RCMs at the entry ports minimizes the possibility of lateral tear resulting from instrument or endoscope motion at the entry surface.

Vision Cart

The vision cart is the central core of the dVSS. It manages the tele-operation logic and processes the video feed captured by the endoscope before sending it out to the

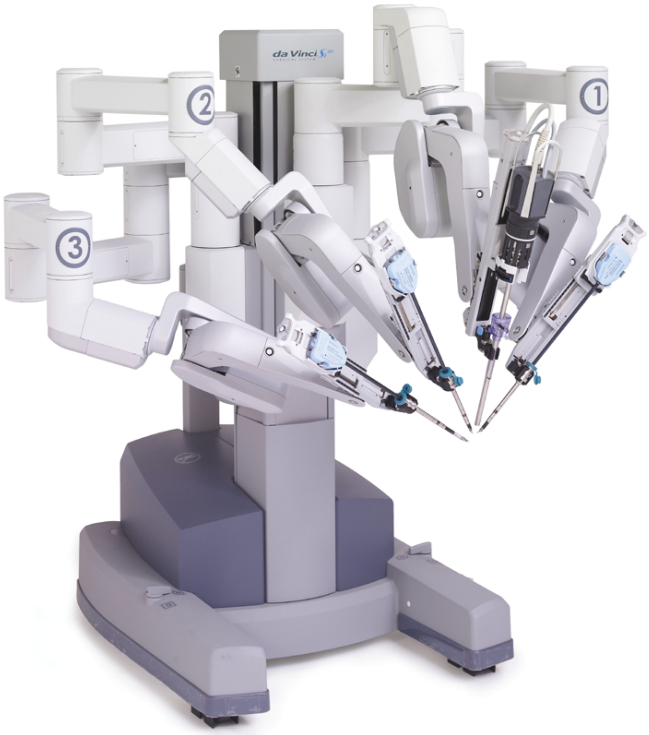


Figure 2.5: Patient Side Cart (PSC) of the dVSS Si version. *Image: © 2017, Intuitive Surgical Inc.*

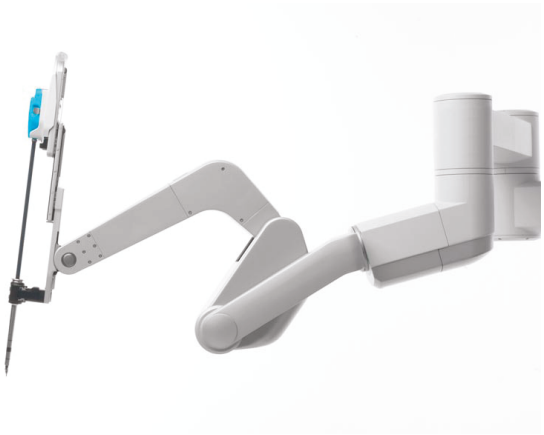


Figure 2.6: Patient Side Manipulator (PSM) of the dVSS Si version. *Image: © 2017, Intuitive Surgical Inc.*

SSC viewer. The vision cart features a touch screen display on top (Figure 2.7) to show the endoscope view to other OR personnel as well as a microphone and speaker to converse with the surgeon sitting at the SSC. A powerful light source generator is present on the vision cart that sends light to the endoscope bulb to illuminate the surgical workspace inside the body. The vision cart hosts an electro surgery unit (ESU) that generates energy for the monopolar and bipolar instruments. ESU gets activated by the foot pedal press on the SSC by the operating surgeon.

EndoWrist® Instruments

The dVSS comes with a large catalog of instruments with a wristed end-effector and a diameter of 8mm. These EndoWrist® instruments bring back the lost dexterity from traditional MIS to RAMIS (Figure 2.8). The surgeon can perform all possible human wrist rotations and more with the capability of these wristed instruments. For example, the instruments allow a roll of more than 180° in either direction which isn't possible with the human hand. The instruments are certified for reuse post standard sterilization processes using autoclave.

In summary, RAMIS using the dVSS maintains the benefits of traditional MIS while bringing back the advantages of open surgery to the surgeon along with higher precision and better instrument control.

Figure 2.7: Vision Cart of the dVSS (picture shown from the Xi model catalog). *Image: © 2017, Intuitive Surgical Inc.*

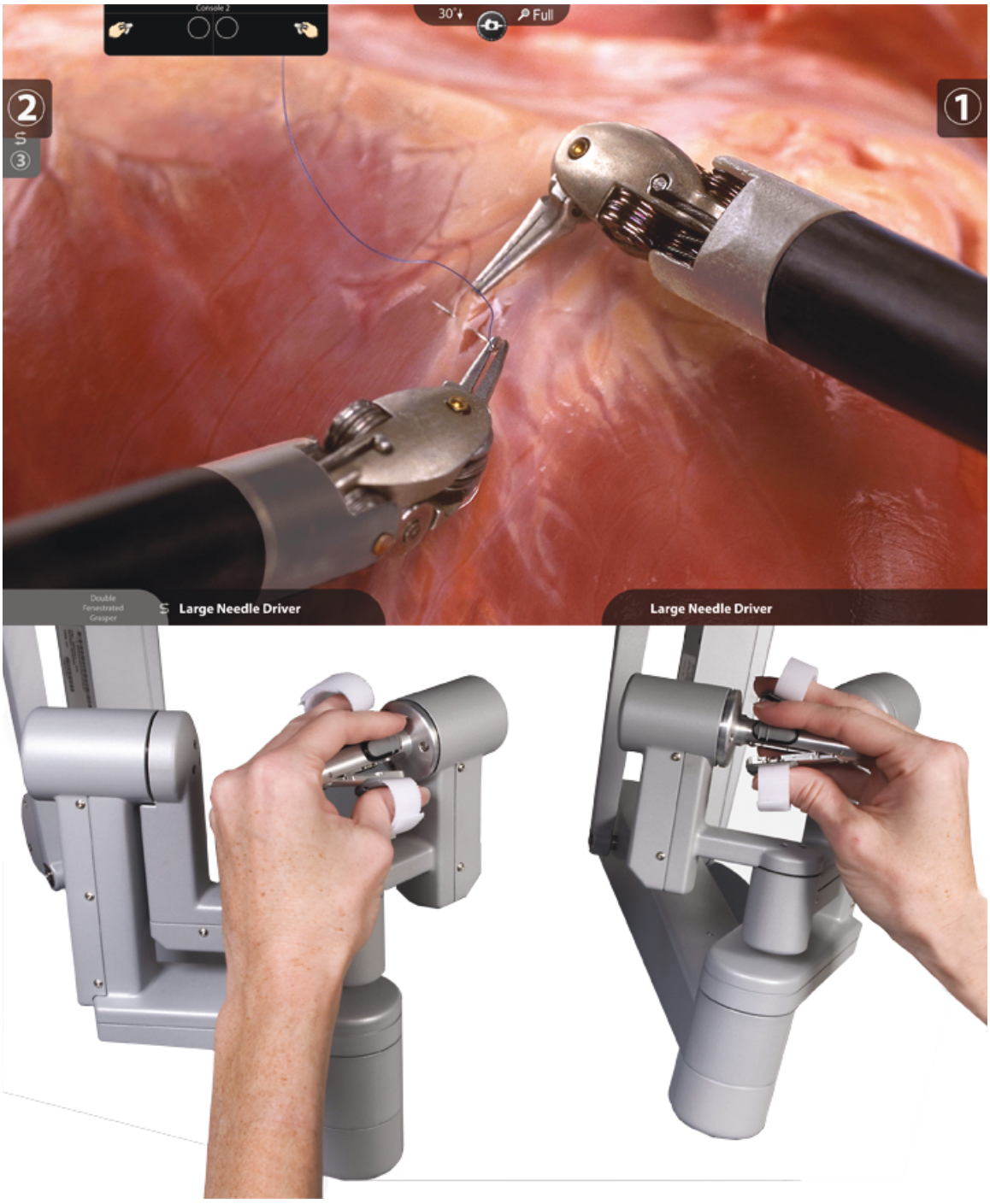


Figure 2.8: Dexterous motion replication from the MTMs (lower half) onto the wristed instrument on the PSMs (upper half) while manipulating a needle with a suture. *Image: © 2017, Intuitive Surgical Inc.*

2.1.2 da Vinci® Skills Simulator™

With limited duty hours,[3] pay-for-performance policies and increasing demand for OR efficiency the resident's OR experience has decreased considerably since 2003.[32] Educators have advocated the use and need for simulation-based training for the development of surgical technical skills of residents.[77] Emphasis on lifelong learning for experienced surgeons with the goal to reduce adverse patient outcomes has increased as well.[95] Simulation-based training can use physical simulation (bench-top models, animal models or cadavers) or virtual reality (VR) simulation (computer graphics and physics modeling engines). Irrespective of the type, simulation provides immediate benefits over OR learning *viz.* flexible scheduling, no risk to patients, learning or developing new techniques – to name a few. Likewise, there are costs associated with simulation training and a thorough analysis accounting for multiple factors[41] needs to be done before choosing the right simulation protocol.

Physical (bench-top) simulation requires more resources compared to VR. Bench-top training tasks require consumables like instruments, tools and physical simulation models that need to be replenished regularly. Laboratory staff has to be present in order to setup, clean up and maintain inventory for efficient training. Physical simulation cannot be available 24/7 without added employment costs for training lab staff. On the contrary, VR simulation does not require such resources and can prove to be an effective solution for learning surgical skills. VR training brings an opportunity to obtain and analyze surgical performance data at scale and with ease,

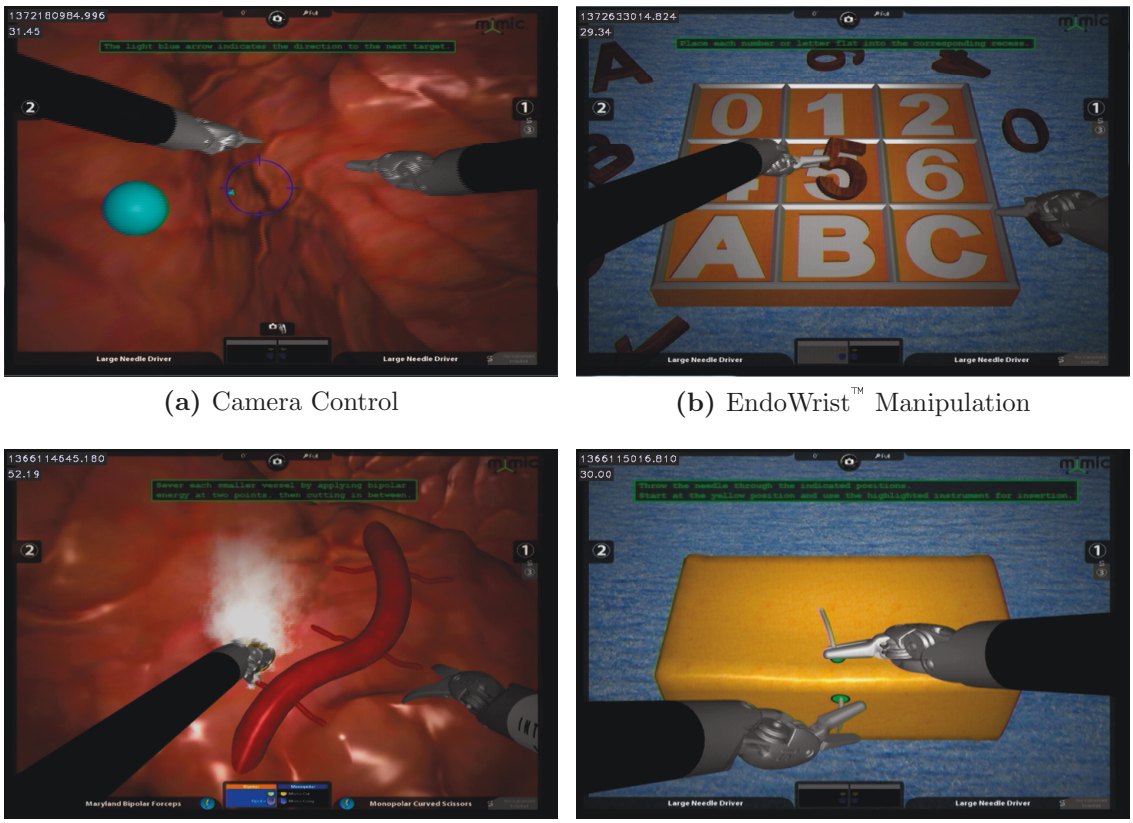
owing to the underlying computer-based architecture. Most VR simulators currently perform such analysis and present evaluations to the trainee. Currently, multiple RAMIS simulators are commercially available.[96,97] Let us look at the da Vinci® *Skills Simulator*™ (dVSim) from Intuitive Surgical Inc..

The dVSS (Section 2.1.1) is a large system in terms of size and cost. It is not financially viable to own a dVSS specifically for training purposes. After-hours use of an existing dVSS in the OR for training might be an option, but definitely not a flexible one. Instead the hospitals can provide a stand-alone SSC with a VR simulator in the form of dVSim for surgical training. This setup requires less space, has no recurring costs (except for annual maintenance), and can become a dedicated training system that is available 24/7. The dVSim is essentially a backpack (Figure 2.9) that hangs on the back of the SSC and connects to it through an optical fiber cable for transferring surgeon's hand motion and simulation video data. It works with the Si version of the dVSS and can be used as a simulator across multiple SSCs owing to its portable design. Internally, the backpack consists of an embedded PC that can render a virtual PSC and perform the functions of vision cart to the extent required for simulation purposes. The SSC continues to function in the usual way except that a 3D rendering of a virtual PSM and endoscope is seen through the stereo viewer. The ability to use the same hardware for simulation and learning as for operating on patients is one of the advantages of the dVSim over the other RAMIS simulators like the Robot Surgical Simulator (RoSS™ ; Simulated Surgical Systems LLC, San Jose,

California) and dV-Trainer™ (dVT; Mimic Technologies, Seattle, Washington). This has been indicated as a plus by trainees in a previous study comparing dVSim and dVT.[98]

The dVSim contains a range of training task modules – from basic system introduction and usage skills to advanced needle driving and suturing skills (Figure 2.10). Some of the technical skills currently included in the dVSim are camera control and targeting, EndoWrist manipulation, blunt dissection, energy-based dissection and transection, needle passing (driving), knot tying and suturing, and advanced instrument (stapler, vessel sealer) usage. A training curriculum using these tasks can be set for trainees based on desired surgical discipline or procedure. The simulation content development is outsourced to two vendors - Mimic Technologies and 3D Systems (formerly Simbionix Ltd.). More recently, Intuitive Surgical Inc. has started developing their own simulation content in collaboration with SenseGraphics AB (Kista, Sweden). The trainee at the end of each training task attempt is shown a performance assessment using the *MScore*™ (Figure 2.11; for modules developed by Mimic only) and the *SimScore*™ (Figure 2.12; Intuitive Surgical Inc., on all the modules). Until very recent, procedure simulations were missing on the dVSim platform. Now, total hysterectomy simulation is available as an add-on to the dVSim platform. Also, the current framework for dVSim does not provide coaching or mentoring, it is just a simulation platform for unstructured and non-individualized training. However, it should be noted that such VR platforms are well suited for the implementation of

Figure 2.9: The da Vinci Skills Simulator is a back pack computer that hangs on the back of the SSC and connects to it using an optical fiber cable. The dVSim simulates the PSC and vision cart using graphics and physics rendering engines. *Image: © 2017, Intuitive Surgical Inc.*



(a) Camera Control **(b)** EndoWrist™ Manipulation

(c) Energy Dissection **(d)** Needle Passing

Figure 2.10: Different surgical technical skills modules available on the dVSim automated surgical coaching.

2.1.3 da Vinci Research API

In 2008, Intuitive Surgical Inc. announced a research interface[99] known as the da Vinci API (Application Programming Interface) to support third party development and collaboration with academic researchers. The da Vinci API provides access to a real-time stream of kinematics of the tele-operation in the dVSS. This includes the two MTMs, three PSMs and ECM. The motion data fields are listed in the Table 2.1. The

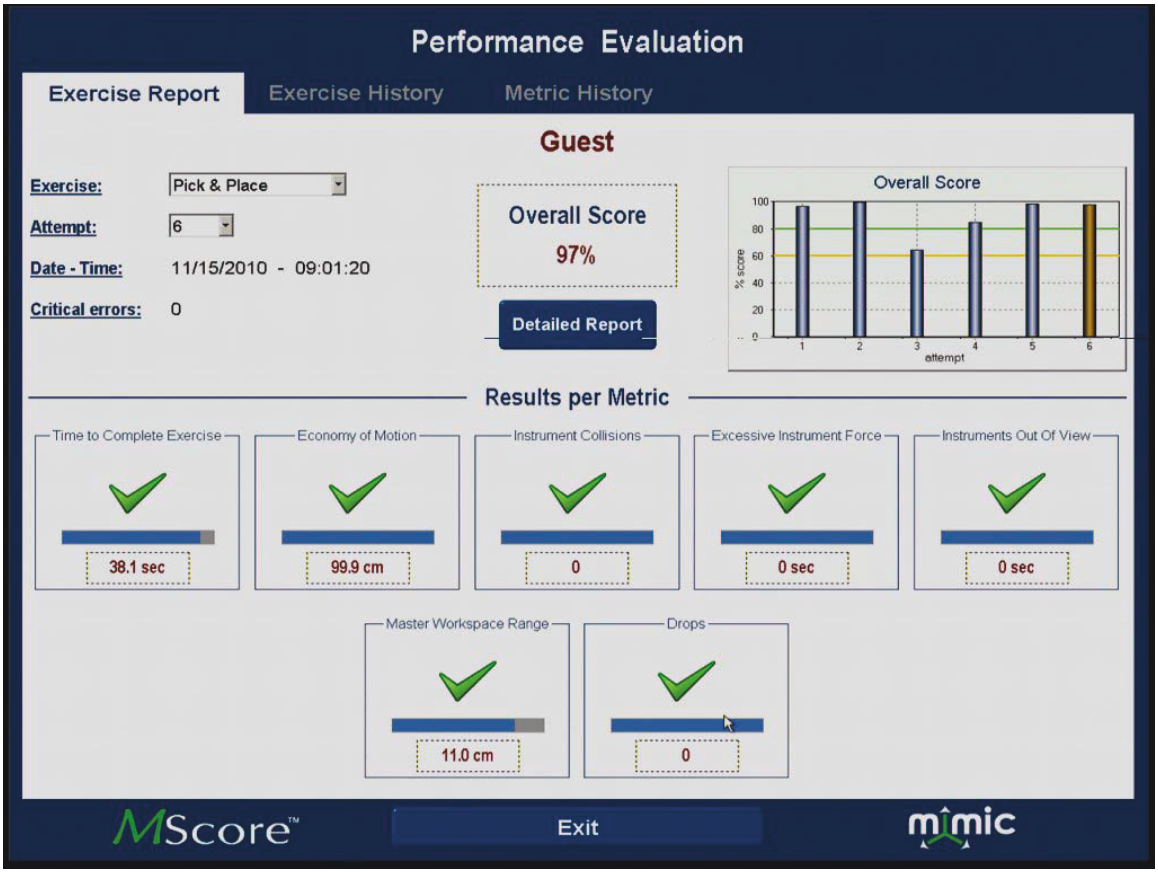


Figure 2.11: A screen capture from the dVSim showing the MScore™ generated after completion of one of the training tasks. The graph on the top right shows a performance trend for the particular user, if any. Various quantitative metrics on motion efficiency and errors accrued are show for the current performance in the center.

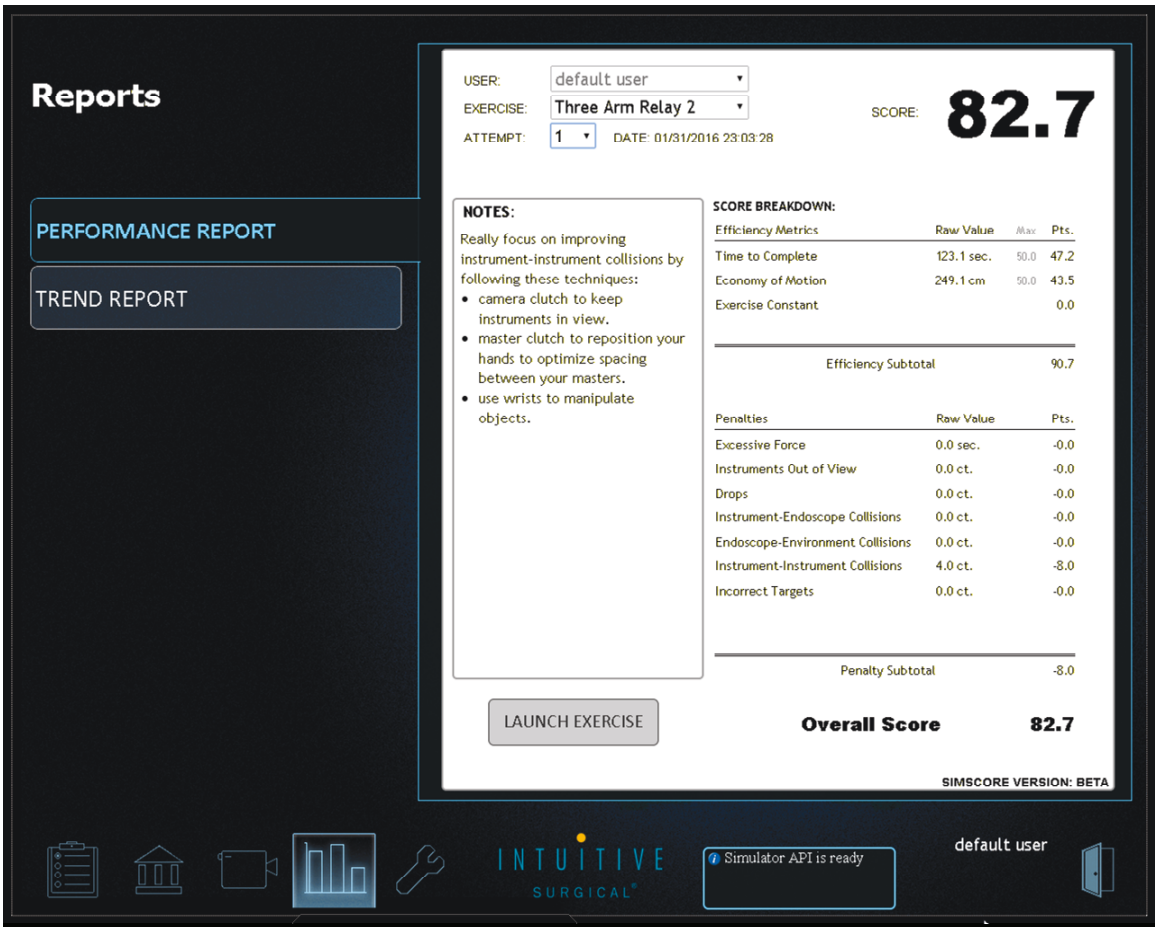


Figure 2.12: A screen capture from the dVSim showing the SimScore™ generated after completion of one of the training tasks. Performance is assessed in terms of motion efficiency metrics as well as errors accrued in the task. A performance trend is available under the other tab of the left navigation sidebar. *Image: © 2017, Intuitive Surgical Inc.*

da Vinci API also provides a real-time signal about any system and user events that occur due to a button or foot pedal press on the dVSS. Some of the key events used later on in the thesis are listed in Table 2.2 with descriptions of the triggering actions. Access to such a combination of hardware-software interface is key to the procurement of large scale data sets. Over the last two decades, the dVSS have spread across the world delivering RAMIS procedures in multiple surgical disciplines and currently over 3200 systems are being used in the ORs around the world. The dVSS (along with the da Vinci API) is a fine example of a CIS device that is slowly becoming an integral part of the OR and making ubiquitous data collection a possibility.

2.1.4 da Vinci Data Recorder

Surgical coaching requires previously collected data for the simple use case of showing demonstrations to trainees from different scenarios and expertise levels. With the introduction of the da Vinci API, non-disruptive data capture of surgical motion and video inside ORs and training labs became easy – at least for the case of RAMIS procedures. The foundation of the National Science Foundation (NSF) Engineering Research Center for Computer-Integrated Surgical Systems and Technology (ERC-CISST) at Johns Hopkins University (JHU) led to a major academic collaboration with Intuitive Surgical Inc. in sharing the da Vinci API. The main objectives of this ERC led to the development of the open source *cisst* libraries[100] and the Surgical Assistant Workstation (SAW) framework.[101] The *cisst* libraries are

Table 2.1: da Vinci API Kinematics Fields

	Field	Description
PSMs and ECM	Instrument Tip Position	3D Cartesian coordinates of instrument tip mounted on the PSM (endoscope tip in case of the ECM)
	Instrument Tip Orientation	3D rotation matrix representing orientation of instrument tip mounted on the PSM (endoscope tip orientation in case of the ECM)
	Instrument Tip Velocity	translational and rotational velocity of instrument tip on the PSM (endoscope tip in case of the ECM)
	Joint Angles	angles / displacements of each PSM / ECM active joint
	Joint Velocity	rotational or translational velocity of each PSM / ECM active joint
MTMs	End-effector Tip Position	3D Cartesian coordinates of the MTM end-effector
	End-effector Tip Orientation	3D rotation matrix representing orientation of the MTM end-effector
	Joint Angles	angles of each MTM joint
	Joint Velocity	rotational velocity of each MTM joint

Table 2.2: da Vinci API Events

Event	Description
MTM Clutch Button	the finger clutch button on either MTM is pulled back or released
Clutch Pedal	foot tray pedal is pressed or released to disengage MTM from the PSM for re-positioning
Camera Control Pedal	foot tray pedal is pressed or released to control the ECM using the MTMs
Head In Sensor	user's head is close to the eyepiece and looking into the stereo viewer or user's head is out
Third Arm Swap	foot tray pedal is pressed or released to start or stop controlling the third PSM
Energy Pedal(s)	foot tray pedal is pressed or released to trigger the electrosurgical instruments using the cut or coagulation energy modes
Instrument Installed	a new instrument is mounted on either of the PSMs
Instrument Removed	a mounted instrument is removed from either of the PSMs

aimed to make software development for computer assisted interventions (CAI) easy. While, the SAW framework combines computer vision, robotics and intra-operative imaging techniques to enhance the surgeon's capabilities. These are available under the open source license at the following URLs: `https://github.com/jhu-cisst` and `https://github.com/jhu-saw` respectively. The SAW framework includes the `sawIntuitiveDaVinci` component that provides a wrapper for the da Vinci API to use it with the rest of the cisst libraries in an easy plug and play format.

The requirements for our projects related to surgical activity recognition, surgical skill assessment and surgical coaching were:

- ✔ ability to record endoscope video, da Vinci API kinematics and events with timestamp-based synchronization,
- ✔ record the data with accuracy and reliability,
- ✔ across different RAMIS training tasks and OR procedures,
- ✔ across different dVSS and at multiple hospital locations,
- ✔ in a non-disruptive manner,
- ✔ with a simple UI for usage by surgeons and operating room staff.

With these in mind, we developed the da Vinci data recorder tool (dVRecorder). The dVRecorder consists of: (1) a computer for storing the collected data, (2) Ethernet and video cables for data logging, and (3) a touch screen (tablet device) with wireless network connectivity to interface with the data collection as shown in the Figure

2.13. The recording computer houses a dual video capture capability. We used two BlackMagicDesign DeckLink Mini Recorder PCI-e cards with SDI input ports. The `cisstStereoVision` library provides wrappers for multiple video capture devices - the DeckLink card was one of them. The recording computer is stowed in one of the storage racks on the vision cart of the dVSS. The computer connects to the Ethernet port on the back of the vision cart to stream da Vinci API data (robot kinematics and events). It connects to the output video ports on the back of vision cart to capture the live endoscope video feed. [1] Thus, no cables go across any portion of the room and the system completely resides on the vision cart. This enables a non-disruptive data collection process. The tablet device connects to the overlay wireless network to access a UI that allows to start and stop the data recording. A basic UI is implemented to make it simple to use by the clinical personnel.

The dVRecorder captures all possible kinematics and events available in the da Vinci API along with stereo endoscopic video as seen by the surgeon on the SSC. The API data is stored in multiple text files – one for each manipulator (2 MTMs, 3 PSMs and 1 ECM) and one for the events log. A universal timestamp using the recording computer's system clock is assigned to each entry of data stored. Video frames are stored in two files with pre-specified file formats and video compression codecs. Each video file is accompanied with a timestamp file mapping video frame number to a universal timestamp based on the same system clock as the API data. Using this

[1] Stereo output ports are add-on and not available by default on the dVSS Si model, in which case the ports on back of SSC may be used for stereo capture.

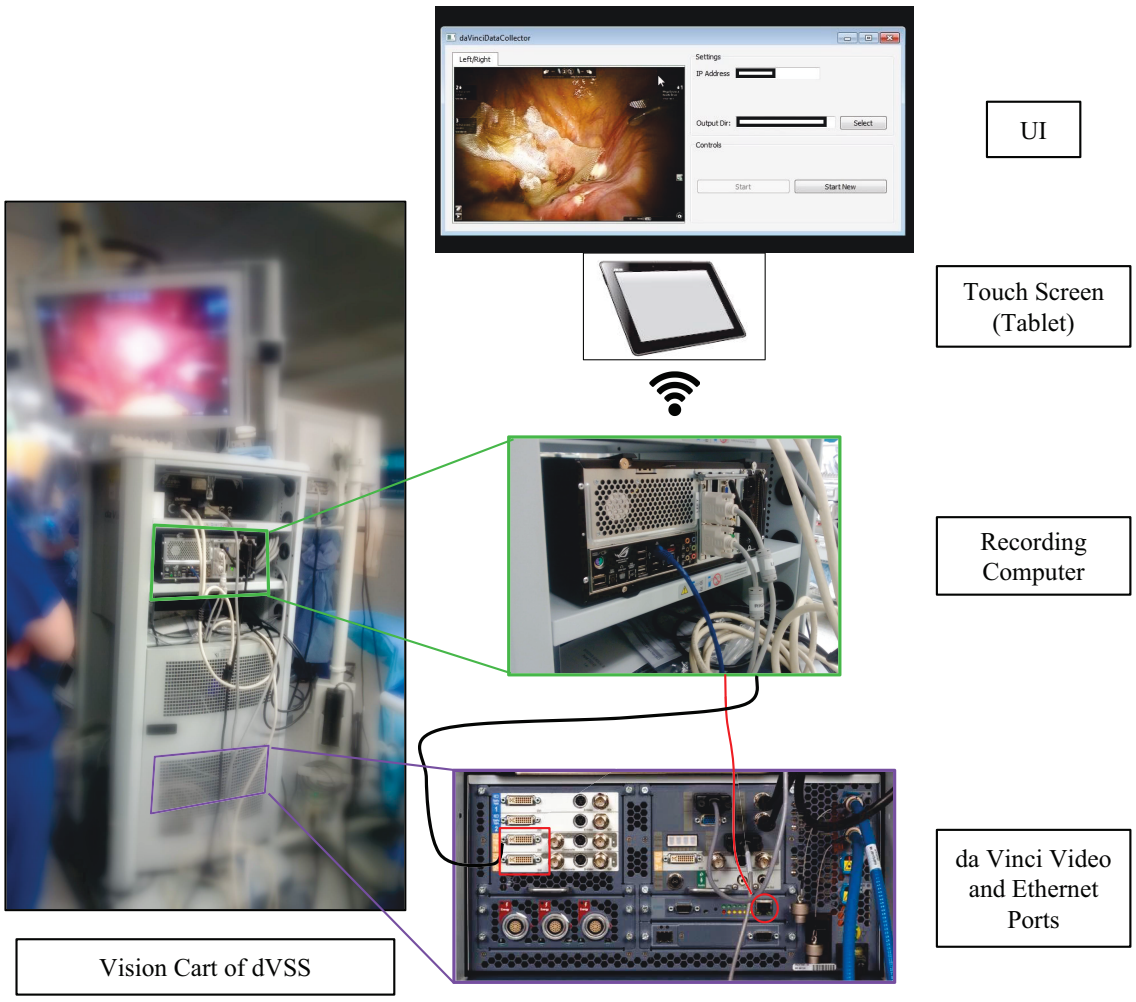


Figure 2.13: da Vinci Data Recorder setup in the OR. The red line (Ethernet cord) connects the Ethernet port of recording computer to that present on the vision cart. The black line (video cables) connects the DVI input ports on the recording computer to the output ports on the vision cart. The tablet device provides a UI for recording controls and sends the commands requested over wireless network to the recording computer.

timestamp, a video frame can be associated to the closest kinematic frame. Note, we refer to the closest kinematic frame, since the two data types - kinematics and video are recorded different framerates (kinematics at 60 Hz compared to the videos at 15 Hz-30 Hz).

A block diagram showing the cisst library and SAW framework components of the dVRecorder is shown in Appendix B along with the challenges faced in OR data recording.

In summary, healthcare providers are increasingly using CIS devices in the ORs making it more and more feasible to record surgical data seamlessly and without interrupting the OR workflow. Surgical data acquisition has multiple hurdles compared to data capture in other fields like computer vision and speech processing. The ethics board (Institutional Review Board) is much more strict about data collection, storage and privacy concerns from the patient and surgeon viewpoints. Introducing a new device (e.g. dVRecorder) requires working with the clinical engineering team to make sure the device complies with the electrical code and policies for surgical care facilities. The operating room technical and nursing staff need to be made aware that the device would stay in the OR for recording purposes. They should be aware of any possible disruptions it could cause in the usual workflow. The entire staff must be onboard for a successful data collection. We have gone through this process multiple times over the past decade to collect rich and interesting data sets from training labs and more recently from the OR. We will present some of these data sets in the next

section.

2.2 Surgical Data Sets

The data sets listed in this section are not yet publicly available. We do intend to release them for non-commercial use in the future. The process of making the data available publicly will require approval from the IRB at JHU as well as from Intuitive Surgical Inc. (since some of the data collected through the da Vinci API is proprietary to ISI). We will first introduce some relevant concepts and then describe the data sets.

Surgical Activity and Task Decomposition

In the past,[102–104] surgical tasks have been decomposed into finer activity segments similar to breaking down of human speech into phonemes. We will explain the activity decomposition and associated terms in brief here, please refer to the papers by Vedula et al.[104,105] for detailed analysis and descriptions of surgical activity segments. We use a hierarchical decomposition of surgical activity – procedures ⇒ tasks ⇒ maneuvers ⇒ gestures. Let us look at examples of activities at these segment levels in the context of a MIS approach for total hysterectomy which is the surgical removal of the uterus and cervix.

- A *procedure* is the entire activity from the start of surgical intervention making

incisions on the belly to access the abdomen cavity to stitching up the incisions at the end.

- A *task* is the activity of closing the vaginal opening created after removal of cervix by a suturing technique.
- A *maneuver* is the sub-task activity of placing a single throw during the knot tying or passing the needle through from one side of the opening to the other.
- A *gesture* is a sub-maneuver activity (and the smallest segment in our framework) of grasping the needle or pulling the suture tail through the wraps.

Surgical Skill Assessment Tools

In the surgical community a variety of assessment scales and tools have been proposed for standardized evaluations of surgical proficiency. We talk about the Objective Structured Assessment of Technical Skills (OSATS)[1] and Global Evaluative Assessment of Robotic Skills (GEARS)[2] here. Both tools involve a video review-based evaluation of skill over certain components or criteria. An evaluator provides their rating on a Likert-like scale of 1 to 5 (1 being poor skill level and 5 being excellent skill level). There are many validation studies on OSATS and relatively smaller number on GEARS (reason being that GEARS is more recent and applicable to just RAMIS procedures). Anchor descriptions for each skill component at the Level 1, 3 and 5 are provided in text on the evaluation forms. A sample form is shown in Figure 2.14 for

OSATS and Figure 2.15 for GEARS.

Compliance with Ethical Standards

We conducted human subjects research protocols to obtain the mentioned data sets. While doing so, all procedures that were performed in research studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. We obtained an informed consent from all individual participants included in the study.

Summary of Data Sets

We have presented an overview of the data sets in the tables that follow. Detailed descriptions of the study protocol, task flow, participant recruitment, data collected and access are presented in Appendix A.

1. MultiSite Suturing Data Set: Table 2.3, Appendix A.1
2. ISI-SG-Sim Needle Passing Data Set: Table 2.4, Appendix A.2
3. WarmUp Hysterectomy Data Set: Table 2.5, Appendix A.3
4. FESS Targeting Data Set: Table 2.6, Appendix A.4

Please circle the number corresponding to the candidate's performance in each category, irrespective of training level.

	1	2	3	4	5
Respect for Tissue:	Frequently used unnecessary force on tissue or caused damage by inappropriate use of instruments		Careful handling of tissue but occasionally caused inadvertent damage		Consistently handled tissues appropriately with minimal damage
Time and Motion:	Many unnecessary moves		Efficient time/motion but some unnecessary moves		Clear economy of movement and maximum efficiency
Instrument Handling:	Repeatedly makes tentative or awkward moves with instruments by inappropriate use of instruments		Competent use of instruments but occasionally appeared stiff or awkward		Fluid moves with instruments and no awkwardness
Knowledge of Instruments:	Frequently asked for wrong instrument or used inappropriate instrument		Knew names of most instruments and used appropriate instrument		Obviously familiar with the instruments and their names
Flow of Operation:	Frequently stopped operating and seemed unsure of the next move		Demonstrated some forward planning with reasonable progression of procedure		Obviously planned course of operation with effortless flow from one move to the next
Instruction of Assistants:	Consistently placed assistants poorly or failed to use assistants		Appropriate use of assistants most of the time		Strategically used assistants to the best advantage at all times
Knowledge of Specific Procedure:	Deficient knowledge. Needed specific instruction at most steps		Know all important steps of operation		Demonstrated familiarity with all aspects of operation

OVERALL ON THIS TASK, SHOULD THE CANDIDATE: **FAIL** **PASS**

Reznick et al. Testing technical skill via an innovative "bench station" examination.
Martin et al. Objective structured assessment of technical skill (OSATS) for surgical residents.

Figure 2.14: Sample form for Objective Structured Assessment of Technical Skills (OSATS) tool[1]

Please circle the number corresponding to the candidate's performance in each category, irrespective of training level.

Depth Perception:	1	2	3	4	5
	Constantly overshoots target, wide swings, slow to correct		Some overshooting or missing of target, but quick to correct		Accurately directs instruments in the correct plane to target
Bimanual Dexterity:	1	2	3	4	5
	Uses only one hand, ignores non-dominant hand, poor coordination		Uses both hands, but does not optimize interactions between hands		Expertly uses both hands in a complementary way to provide best exposure
Efficiency:	1	2	3	4	5
	Inefficient efforts; many uncertain movements; constantly changing focus or persisting without progress		Slow, but planned movements are reasonably organized		Confident, efficient and safe conduct, maintains focus on task, fluid progression
Force Sensitivity:	1	2	3	4	5
	Rough moves, tears tissue, injures nearby structures, poor control, frequent suture breakage		Handles tissues reasonably well, minor trauma to adjacent tissue, rare suture breakage		Applies appropriate tension, negligible injury to adjacent structures, no suture breakage
Autonomy:	1	2	3	4	5
	Unable to complete entire task, even with verbal guidance		Able to complete task safely with moderate guidance.		Able to complete task independently without prompting.
Robotic Control:	1	2	3	4	5
	Consistently does not optimize view, hand position, or repeated collisions even with guidance		View is sometimes not optimal. Occasionally needs to relocate arms. Occasional collisions and obstruction of assistant.		Controls camera and hand position optimally and independently. Minimal collisions or obstruction of assistant
Use of Third Arm:	1	2	3	4	5
	Consistently does not use it, or does not use it well when required, even with verbal guidance.		Mostly uses 3rd arm in a safe and efficient manner with moderate guidance.		Consistently uses 3rd arm in a safe and efficient manner without prompting.

Goh et al. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills.

Figure 2.15: Sample form for Global Evaluative Assessment of Robotic Skills (GEARS) tool[2]

Table 2.3: MultiSite Suturing Data Set (Appendix A.1)

Name	MultiSite
Surgery Type	Robot-assisted Minimally Invasive Surgery (RAMIS)
Instruments	da Vinci Surgical System (dVSS)
Task	Bench-top interrupted suturing
Participants	14 surgical residents and 4 attending surgeons
Size	135 repetitions
Data Collected	da Vinci API kinematics and events, stereo endoscope video
Data Annotated	single expert OSATS scores, framestamp-based activity labels (gestures and maneuvers)

Table 2.4: ISI-SG-Sim Needle Passing Data Set (Appendix A.2)

Name	ISI-SG-Sim Needle Passing
Surgery Type	Robot-assisted Minimally Invasive Surgery (RAMIS)
Instruments	da Vinci Skills Simulator (dVSim)
Task	Virtual reality needle passing
Participants	24 engineers and 6 RAMIS trainers
Size	102(+42) repetitions
Data Collected	da Vinci kinematics and events, simulation events, simulation kinematics, task progress log, mono endoscope video
Data Annotated	automated maneuvers segments

Table 2.5: WarmUp Hysterectomy Data Set (Appendix A.3)

Name	WarmUp Hysterectomy
Surgery Type	Robot-assisted Minimally Invasive Surgery (RAMIS)
Instruments	da Vinci Surgical System (dVSS)
Task	Live robot-assisted laparoscopic hysterectomy procedures on patients
Participants	23 residents and 6 attending gynecologists
Size	27 procedure recordings
Data Collected	da Vinci kinematics and events, stereo endoscope video
Data Annotated	OSATS and GEARS ratings from operating attending surgeon and resident, task segments labels with timestamps

Table 2.6: FESS Targeting Data Set (Appendix A.4)

Name	FESS Targeting
Surgery Type	Endoscopic Surgery
Instruments	Nasal Pointer and Endoscope
Task	Cadaver head-based simulation
Participants	13 residents and 7 attending surgeons
Size	49 session recordings
Data Collected	instrument and endoscope position and orientation, surgeon's eye-gaze location, mono endoscope video
Data Annotated	binary and Likert-like (1 to 5 ordinal) ratings from three individuals, task segment labels with timestamps

2.3 Performance Metrics

Need for objective and automated skill assessment has been advocated by surgical educators.[77] Currently, assessment tools like OSATS and GEARS are the gold standard for technical skill assessment. However, these tools require significant effort and time from experienced surgeons or instructors to review videos of task performances and grade them. Such reviews are still subjective and biased by reviewer's preference of operating styles. Thus, a truly objective assessment might be possible if a pool of experts independently review the same performance. Although, this just increases the amount of time required collectively from attending surgeons which is already a scarce resource due to operating schedules, clinic duties and administrative work. Evaluating a surgeon's technical skills using motion data from the performance has shown success and validation using a variety of approaches.[53,54,60,106–111]

In this section, we will define some of the motion data-based performance metrics that have been previously used in the domain of surgical skill assessment. We will use these metrics in the later chapters of this thesis. We assume that motion data is always accompanied with a timestamp or some similar attribute related to the notion of time (e.g. framestamp). Let us define some notation before we define and formulate the various metrics. Let the data stream be indexed using i. Let us denote timestamp using t. Let 3D position be the row vector, $\mathbf{p} = [x\, y\, z]$, and 3D orientation be the row vector $\mathbf{o} = [x\, y\, z\, \theta]$ in the axis angle notation. The orientation may be represented using a 3×3 rotation matrix notation as $\mathbf{R}$ with $\mathbf{R}_x$, $\mathbf{R}_y$ and $\mathbf{R}_z$ as the

rows. Also, let us assume that the segment of interest in the data stream starts at index $i = 0$ and ends at $i = T$.

Completion Time

This is the total time elapsed from start to end of the task performance computed as follows:

$$CT = t_T - t_0 \tag{2.1}$$

Path Length

This is the total length of the path traced by the position data from start to end of the task. This can be computed as follows:

$$PL = \sum_{i=0}^{T-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2} \tag{2.2}$$

Ribbon Area

This metric is specifically defined for wristed instruments.[112] It is the total area swept by the vector joining the base and center of the wrist as shown in Figure 2.16. Let the center of wrist be denoted by p and base by q. Then, ribbon area metric is computed as follows:

$$RA = \sum_{i=0}^{T-1} A(p_i, q_i, p_{i+1}, q_{i+1}) \tag{2.3}$$

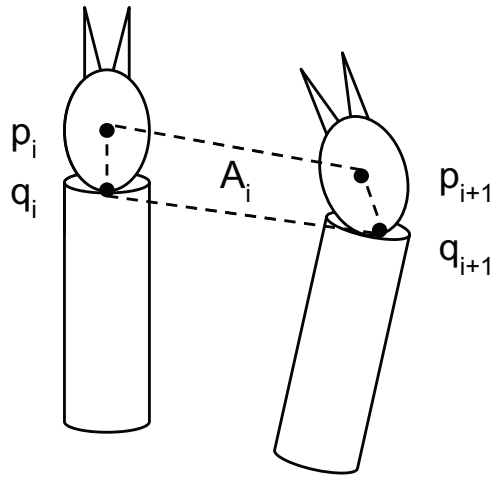


Figure 2.16: A cartoon sketch showing ribbon area metric when instrument moves between data index i and $i+1$. It is the area of the quadrilateral formed by the points p_i, q_i, p_{i+1} and q_{i+1}.

where, A is the area of the quadrilateral formed by the points p_i, q_i, p_{i+1} and q_{i+1}.

Movements

A *movement* is defined as a peak in the magnitude of velocity of instrument motion. For this, we first compute the velocity magnitude v as follows:

$$v_i = \sqrt{\left(\frac{x_i - x_{i-1}}{t_i - t_{i-1}}\right)^2 + \left(\frac{y_i - y_{i-1}}{t_i - t_{i-1}}\right)^2 + \left(\frac{z_i - z_{i-1}}{t_i - t_{i-1}}\right)^2} \tag{2.4}$$

We apply a 1-D median filter with a window size equivalent to a time span of one third of a second to filter high frequency noise present. Previous works have performed empirical analysis to determine a threshold for filtering such high frequency noise.[53] The filtered velocity magnitude is processed to identify and count the number of

peaks in given data, Δ. Finally, we compute the metric as number of movements per second (time quantization):

$$MV = \frac{\Delta}{t_T - t_0} \tag{2.5}$$

Gripper Activations

This is the total number of times the gripper (jaw) of the instrument is closed from start to end of the data. Given the joint value for gripper as j also indexed using i and a threshold $thresh$, we first compute a binary mask as follows:

$$m_i = \begin{cases} 1 & j_i \leq thresh \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

We, then, compute the metric as number of times the value of m goes from 0 to 1:

$$GA = \sum_{i=0}^{T} gt(m_{i+1} - m_i, 0) \tag{2.7}$$

where, $gt(\cdot)$ is a comparator function on the $>$ operator and gives a binary output of 0 or 1. Ideally, $thresh$ should be 0 for a closed gripper, However, in reality, an empirical value is computed by experimentation for $thresh$.

Master Workspace Volume

This is the volume in 3D space within which motion of surgeon's hands (MTMs in case of RAMIS using dVSS or dVSim) is contained. We compute this as volume of the convex hull encompassing the 3D point cloud formed by data from start to end of performance as follows:

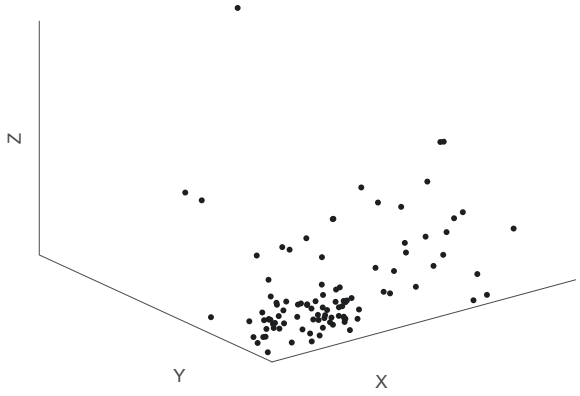
$$MWV = volume(conv\_hull(p_0, p_1, \ldots, p_T)) \tag{2.8}$$

where, *volume* computes the volume of a convex hull. We use MATLAB's `convhulln` method and Python-based sci-py package `scipy.spatial.ConvexHull` class to compute the volumes. Figure 2.17 shows a simulated point cloud and the fitted convex hull and its volume.

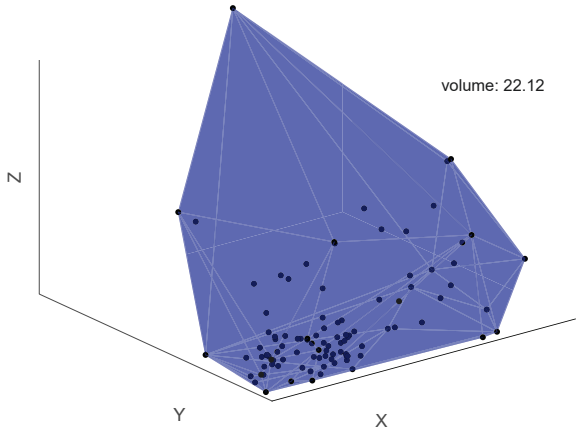
In addition to the above performance metrics, we do compute some specialized metrics for evaluating skill on a data set specific basis. These specialized metrics may also be computed because of additional data fields available in the particular data set. We will define these metrics as and when needed in the later chapters of the thesis.

2.4 Crowdsourcing

As introduced earlier, the collective wisdom of humans (crowds) with certain characteristics may be good at providing solutions for tasks that are not within the reach



(a) point cloud



(b) convex hull and volume

Figure 2.17: Master Workspace Volume metric

of computers and algorithms yet.[113] A task may be challenging due to the time or space complexity associated with it e.g. folding and building protein structures.[17] In other case, the task could be to generate reliable ground truth data to train computer algorithms from machine learning and pattern recognition.[28,114,115] In this thesis, we investigate crowdsourcing-based user studies to analyze the ability of the crowd to summarize surgical data. The core concepts that introduce the terminology used in the remainder of this thesis with respect to crowdsourcing are described in detail in Appendix C. We introduce the core concepts related to crowdsourcing and explain them below, since the novel use cases in our studies are spread across multiple chapters.

Definition

The authors Estellés-Arolas and González-Ladrón-de-Guevara in[116] built an integrated definition for crowdsourcing by surveying milestone works in the domain of crowdsourcing. The definition is quoted from the paper below:

> Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken.

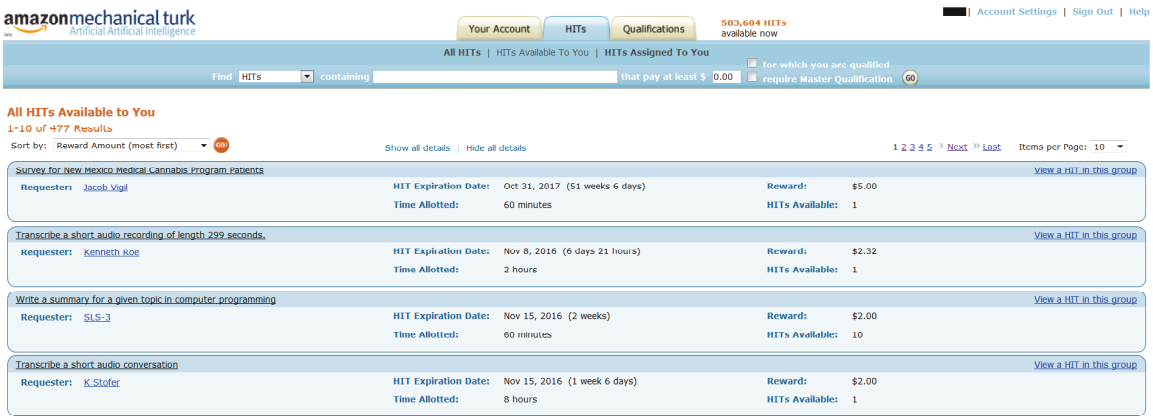


Figure 2.18: A screen capture showing the MTurk portal for crowdsourcing tasks.

Amazon Mechanical Turk (MTurk)

For most of our work with crowdsourcing we have utilized the web-based platform "Mechanical Turk" created and maintained by Amazon.com, Inc. (Seattle, Washington) `https://www.mturk.com/`. MTurk is an online marketplace for undertaking crowdsourcing work (a screen capture is shown in Figure 2.18). We will explain some of the concepts useful for our work below. Please refer to the FAQ section (`https://requester.mturk.com/help/faq`) or MTurk's website for more details.

2.5 Inter-rater Agreement

While validity of the crowdsourcing is important, the other important test for success of crowdsourcing is measure of reliability within crowd compared to that within experts. Inter-rater agreement (reliability) gives us a measurement of consensus between multiple raters. We introduce some measurements for inter-rater agreement

below that are used to test reliability of the crowd in later chapters. We will restrict the discussion to nominal ratings as our HITs contain only nominal-scale categorical questions.

Percent Agreement

Traditionally, inter-rater agreement has been measured simply as percent agreement. The basic idea is that for each task, multiple raters have provided their responses. In case of nominal-scale ratings, we simply compute the fraction of raters who selected the majority (mode) response per task. And, the percent agreement is computed as the average of these fractions over all the tasks. For example, given n tasks and k categories let us assume we can represent the responses in a matrix $\mathbf{R}$ of dimension $n \times k$. The entry R_{ij} is the number of raters who selected category j for the task i, where $1 \leq i \leq n, 1 \leq j \leq k$. We can compute the fraction agreement f_i for task i as:

$$f_i = \frac{max(R_{i,})}{\sum_{j=1}^{k} R_{ij}} \tag{2.9}$$

The percent agreement is computed as:

$$\text{perc agr} = \frac{\sum_{i=1}^{n} f_i}{n} \tag{2.10}$$

kappa Statistic

The underlying assumption that percent agreement makes about raters and responses is that majority raters are correct and are making a thoughtful choice. However, it is possible that the raters were making random guesses at some if not all tasks. Thus, the agreement that we observe by computing percent agreement may not be the true underlying agreement within the raters. For example, given a binary rating task say 70 out of 100 raters selected 'Category A' versus 'Category B'. This results in a 70% (0.7) percent agreement. Now, if a random number generator was used to perform this rating, it is possible that 50 times out of 100 it would select A versus B resulting in a 50% percent agreement. Thus, a measurement of inter-rater agreement should account for this chance agreement. This is what was proposed by Cohen[117] in 1960 when he defined Cohen's kappa as a statistic to measure inter-rater agreement.

Cohen's kappa

Cohen's kappa (κ) is defined for nominal-scale ratings obtained from two raters over a fixed number of tasks and categories. Assuming similar notation from before, let there be n tasks and k categories, we can obtain a matrix $\mathbf{R}$ of size $k \times k$. The rows are for rater 1 and columns for rater 2 such that, R_{ij} indicates the number of tasks for which rater 1 assigned category i and rater 2 assigned category j to the same task. Now, convert the response matrix to a proportion matrix, P such that

$P_{ij} = R_{ij}/n$. Then, observed agreement can be computed as p_o,

$$p_o = \sum_{i=1}^{k} P_{ii} \tag{2.11}$$

The expected agreement by chance can be computed as p_e,

$$p_e = \sum_{i=1}^{k} P_{i,}.P_{,i} \tag{2.12}$$

Finally, κ can be computed as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2.13}$$

Please refer to Section 18.1 in Fleiss et al.[118] for explanation and derivation of these equations with examples. Also, refer to Equation 18.13 for computing the standard error of κ. A value of $\kappa = 1$ indicates perfect agreement and $\kappa \leq 0$ indicates no agreement or chance agreement among raters.

Fleiss' kappa

Cohen's kappa was limited to two raters performing fixed number of tasks over fixed categories. In 1971, Fleiss[119] defined a new kappa for inter-rater agreement that generalized to multiple raters with fixed number of raters per task. A version of Fleiss' kappa is also available for the case wherein ratings from different number of raters

per task are obtained. Please refer to Section 18.2 and 18.3 in[118] and Equations 18.44 and 18.48 specifically to understand the computation of the kappa statistic. A value of $\kappa = 1$ indicates perfect agreement and $\kappa \leq 0$ indicates no agreement or chance agreement among raters.

2.6 Summary

In this chapter, we described open surgery, MIS (minimally invasive surgery), and RAMIS (robot-assisted minimally surgery). We introduced the da Vinci® Surgical System (dVSS) and da Vinci® Skills Simulator™ (dVSim) that are currently the most popular systems available to perform RAMIS. We described the da Vinci API and the da Vinci Data Recorder tool which have been instrumental in collecting data from the dVSS and dVSim systems in training labs and operating rooms likewise. We, then, presented a summary of four surgical data sets – MultiSite Interrupted Suturing (bench-top simulation), ISI-SG-Sim Needle Passing (virtual reality simulation), WarmUp Hysterectomy (OR patient) and FESS Targeting (cadaver simulation) data sets. We defined some of the common performance metrics computed using surgical motion (kinematics) data analysis. We introduced a formal definition for crowdsourcing and introduced Amazon Mechanical Turk. Finally, we listed the different inter-rater agreement metrics and their use case from the world of statistics. By doing so, we have laid out the key concepts, data and terminology that will be used and

referred to, through the remainder of this thesis.

Chapter 3

Surgical Activity Modeling

Table 3.1: Activity modeling is an important component required for a surgical coaching framework.

Training Type	Activity Recognition	Skill Assessment	Feedback	Demonstration
VR	✔	🎓[4,5]	🎓[5]	🎓[5]
Bench Top	✔	🎓[4]	✖	✔
OR	✔ 🎓[3]	✔	✖	✖

✔: indicates previous work in literature exists, 🎓[X]: indicates a solution is presented in Chapter "X" of this thesis, ✖: indicates no prior work exists and none is presented in this thesis.

We believe that the key coaching activities (Section 1.3) inherently rely and assume

Work from this chapter has been previously published in a peer-reviewed journal article. A. Malpani, C. Lea, C. C. G. Chen, and G. D. Hager, System events: readily accessible features for surgical phase detection, Int J CARS, vol. 11, no. 6, pp. 12011209, May 2016.[120] The core theory and hypotheses behind the paper's framework, data set generation and cleanup along with the testing of linear SVM and random forests was performed by me (A. Malpani). C. Lea was responsible for the temporal convolutional neural network implementation as well as the segmental inference component of the framework. C. C. G. Chen was the clinical advisor and principal investigator on the study that generated the hysterectomy data set. G. D. Hager was the technical advisor and mentor for A. Malpani and C. Lea.

on the coach's capability to break down trainee performance into relevant segments to evaluate, critique and recommend practice upon. Similarly, the coach should have an indexed and segmented data archive of performances to demonstrate the relevant skill segment to the trainee. While this capability can be assumed in the case of manual coaching by a human, an automated coach requires underlying methods to generate such task breakdowns and answer the trainee's question – *how do I do it correctly?*, *where was I wrong?*, and *what do I do to improve?*. Of course, doing so manually is not feasible at the scale and volume of surgical procedures and training happening around the world and automated solutions are needed. Motivated by this, we present an approach to perform activity modeling to segment the surgical procedure at hand in this chapter.

3.1 Background

Before we present our framework to perform automated surgical activity modeling, we will spend some time to explain (1) the different sets of problems in the domain of activity modeling, (2) the structure in surgical activity, (3) the various data modalities available in surgical scenarios that are relevant to activity modeling, (4) other applications of activity modeling, and (5) prior work in the literature.

3.1.1 Activity Modeling Problems

We use the following terms to describe the different problems in the context of activity modeling.

Classification: This is the process of assigning a single label from a set of activity classes to a given activity segment (pre-specified start and end times).

Segmentation: This is the process of finding all time points during the activity that are boundaries (start and end) of constituent segments. This does not involve assigning a label for each segment.

Recognition: This is the process of joint segmentation and classification i.e. finding the boundary points as well as assigning labels for the constituent segments.

Detection: This is the process of retrieving all segment occurrences (i.e. boundaries) in the activity sequence given a class label.

3.1.2 Structure in Surgical Activity

The Language of Surgery (LOS) project at Johns Hopkins University has a simple hypothesis – surgical activity has an underlying structure.[61,102,103] The motivation for this comes from the speech and language community. When a human talks at length or writes a paragraph, they have a notion about the underlying structure which is governed by rules of the language they are speaking or writing. The grammar and vocabulary form two important components of these underlying rules. A person with

fluency in speaking understands these two components and can construct meaningful sentences using them, with the words forming a structured sequence. If another person was asked to convey the same notion given the list of words, there will be some variation, but not too much, in the sequence they come up with. The factors resulting in this variation could be from the past - social and educational background, and from the present - emotional state, environment and audience. Analogous to this, when surgeons go in the OR to perform a surgical intervention, they understand the anatomy and procedure. They are aware of the steps to be performed and the allowed sequences (grammar) in which they should be performed. This relates to medical school and residency institution the surgeon attended and forms the background (past) factors. As they walk in the OR, they have a new patient (with different anatomy), may be some new staff and assistants, and these form the present factors. The surgical procedure is performed and the activity sequence is dependent on at least these factors. Given another surgeon performing the same procedure, these factors differ and we may see a different sequence.

This led to the development of structuring surgical activity and a hierarchical decomposition of surgical activity as described in Section 2.2 – procedure $\Rightarrow$ phase / task $\Rightarrow$ maneuver $\Rightarrow$ gesture – was constructed. Figure 3.1 shows this hierarchical breakdown of surgical activity in the context of suturing and knot tying. This is also analogous to the speech community where the acoustic signal is broken down from sentence $\Rightarrow$ words $\Rightarrow$ phones $\Rightarrow$ phonemes to perform various speech processing

applications.

3.1.3 Data Modalities

In Chapters 1 and 2, we mentioned how recording intra-operative surgical data from training and OR is becoming increasingly easier with the introduction of more CIS devices. While surgical activity can be structured and decomposed into segments, information conveying this structure can be spread across different sensors in these CIS devices. The three main data modalities that can be obtained from surgical interventions are (1) video (2) motion and (3) other sensors. The video can be capturing the OR workflow or the surgical performance and can be multiview (multiple cameras installed at different position in the room) as well as using RGB-D sensors like the Kinect (Microsoft Corp.). We will refer to the surgical performance video i.e. the view obtained from an endoscope in the rest of our work. Motion data can contain instrument kinematics, endoscope kinematics as well as console kinematics (in case of a tele-surgical system like dVSS). Additionally, there are multiple devices present in an OR that can be sensorized *viz.* cautery tools, suction and irrigation devices, surgical lamps and lights, OR table inclination and so on. While it is not obvious how to obtain such sensor data, we believe that with newer technology this will be possible. For example, we can obtain this information using the da Vinci API (Section 2.1.3). Table 3.2 compares pros and cons of these data modalities with respect to ease of collection, information contained and processing. We believe that differ-

Figure 3.1: The structure in surgical activity and its hierarchical decomposition. This figure appears in Vedula et al.[104]

Table 3.2: Surgical data modalities for activity modeling

Data Modality	Pros	Cons
Video	✔ contains all scene context	✖ challenging to extract context using existing computer vision techniques ✖ privacy and anonymity concerns
Motion	✔ captures intent and planning	✖ lacks scene context ✖ requires sophisticated instrumentation
Sensor	✔ captures some context ✔ easy and cheaper to capture ✔ simple to process (binary)	✖ lacks scene context ✖ lacks intent and planning

ent attributes of surgical activity information are captured across the different data modalities and a joint analysis of these data modalities will lead to highly accurate activity recognition performance.

3.1.4 Significance of Phase Recognition

Recently, surgical procedure assessment has become important with the pay-for-performance policies. In a large clinical study, Birkmeyer et al.[121] showed that post-operative outcomes are associated with technical skills of the operating surgeon. They recommended that peer review may be useful to assess surgical skills. But, such peer review is impractical at scale due to time and resource constraints and may require an organized effort. As a means to achieve this, companies providing commercial

solutions for surgical procedure performance review are popping up like CSATS Inc. (Seattle, Washington) and Sigma Surgical Corp. (Toronto, Canada). While they provide valid and constructive reviews at the phase level using performance videos, they are currently limited from scaling up due to the manual effort required in labeling of phase segments in these videos.

In addition to providing review, such activity segmentation can be used to show relevant pre-segmented performance videos to residents in training as well as experienced surgeons adopting a new operating technique or surgical procedure in their practice. Surgeons can be provided statistics on the phases from their previous procedures along with patient outcomes. Real-time surgical phase recognition can help improve OR and hospital workflow efficiency. For example, staff in pre-operative room can prep the patient for surgery at the right time by relying on a system showing estimated completion times for the current surgery instead of doing it early or late.[122] Such real-time phase recognition can be used to advance the current OR to what is being called as "context-aware OR". Based on the phase information, CIS technologies can display useful data or perform contextual interventions related to the surgical phase.

3.1.5 Previous Work

Recognizing the potential impact and significance of surgical activity modeling, lot of research has been undertaken over the past decade. A variety of techniques

have been developed and validated using approaches from machine learning (including deep learning). We shall talk about these works in brief and organize this review of prior work based on the surgery type i.e. VR simulation, bench-top simulation and OR procedures.

VR Simulation

Performing surgical activity recognition in VR is like trying to solve a problem whose solution is potentially already known. This is well reflected by the absence of previous literature modeling activity just in VR data setting. By virtue of computer simulation, the groundtruth information about scene context, user's intent and planning, as well as interactions with the environment are known de facto. As we show later in Chapter 6, a task progress manager can be developed to track the surgical activity being performed with complete confidence. Of course, VR provides an excellent opportunity to test algorithms developed for bench-top and OR data with readily available training and groundtruth data.

Bench-top Simulation

While the major focus of contributions in this chapter is at the surgical phase level activity, we believe methods targeted at finer grain activity segments like gestures and maneuvers are relevant to the bigger picture of an automated coach. We have listed the previous works on fine-grain activity modeling in bench-top training tasks in Table

3.3.

OR Procedures

Table 3.4 lists some of the previous works grouped by the data modality that have been developed to perform phase activity modeling. While the bench-top work has focused mainly on kinematics and video based approaches, methods for surgical procedure segmentation into constituent phases have used sensor data traditionally owing to the complexity of intra-operative video. Recently, video-based methods have shown state-of-the-art performance on standardized data sets using convolutional neural network approaches.[135–137]

Current Limitations and Solutions

Validation using Laparoscopic Cholecystectomy

All of the above approaches for surgical phase recognition (except the ones using activity triplets in Table 3.4) have been validated on data sets containing laparoscopic cholecystectomy (lap-chole) procedures. Lap-chole is a straightforward surgical procedure with an almost sequential phase flow. Methods validated using such data are prone to learning the sequence of steps and may not show similar validity on data sets containing complex procedures like hysterectomy or prostatectomy. Performance validation on such complex procedures is missing in literature. Additionally, the sensor data-based methods rely on sources that require additional instrumentation of

Table 3.3: Prior work on surgical activity modeling in bench-top simulation

Data	Authors	Task	Method
Kinematics	Lin et al.[102,123]	g.r.	LDA+HMM, PCA+HMM
	Varadarajan[103]	g.r.	FA-HMM, LDS, S-LDS
	Tao et al.[56]	g.c.	Sparse-HMM
	Jun et al.[124]	g.c.	decision trees
	Ahmidi et al.[60]	g.c., g.r.	DCC-CSM
	Despinoy et al.[125]	g.r.	signal processing
	Gao et al.[126]	m.s.	SDAE
	Gao et al.[13]	m.d.	SDAE + AS-DTW
	DiPietro et al.[127]	g.r.	LSTM-RNN
Video	Haro et al.[128]	g.c.	multiple kernel learning
	Lea et al.[129]	g.r.	LC-SC-CRF
	Lea et al.[130]	g.r.	spatiotemporal CNN
	Rupprecht et al.[131]	g.r.	CNN + LC-SC-CRF
	Lea et al.[10]	g.r.	TCN
Kinematics + Video	Zappella et al.[132]	g.r.	multiple kernel learning
	Tao et al.[133]	g.r.	MsM-CRF
	Lea et al.[134]	g.r.	SC-CRF

g.r.: gesture recognition, g.c.: gesture classification, m.s.: maneuver segmentation, m.d.: maneuver detection
LDA: linear discriminant analysis, PCA: principal component analysis, HMM: hidden Markov model, FA-HMM: factor analyzed HMM, LDS: linear dynamical systems, S-LDS: switching LDS, DCC-CSM: descriptive curve coding common string model, SDAE: stacked de-noising auto encoder, AS-DTW: asymmetric subsequence dynamic time warping, CRF: conditional random field, SC-CRF: skip chain CRF, LC-SC-CRF: latent convolutional SC-CRF, CNN: convolutional neural network, TCN: temporal convolutional network

Table 3.4: Prior work on surgical phase recognition in OR procedures.

Data	Authors	Method
Video	†Twinanda et al.[138]	BoW + FK
	Dergachyova et al.[139]	AdaBoost + HsMM
	Twinanda et al.[11]	CNN
	Jin et al.[135]	RCN
	Cadène et al.[136]	CNN + HMM
	Twinanda et al.[137]	CNN + LSTM
Video + Sensor	Blum et al.[140]	CCA + DTW
Sensor	Ahmadi et al.[141]	DTW
	Padoy et al.[142]	AdaBoost + DTW
	Padoy et al.[143]	DTW + HMM
	Stauder et al.[144]	Random Forests
	DiPietro et al.[145]	CRF
Activity Triplets	Katić et al.[146–148]	Ontology-based
	Forestier et al.[149]	Decision Trees

† is a phase detection method, while others are phase recognition.
BoW: bag of words, FK: fischer kernel, HMM: hidden Markov model, HsMM: hidden semi-Markov model, CNN: convolutional neural network, RCN: recurrent convolutional network, LSTM: long short term memory, DTW: dynamic time warping, CRF: conditional random field.

the OR like abdominal CO_2 pressure, irrigation-suction bag weight. This limits the scalability of these approaches, at least at current times.

We will present our surgical phase recognition framework in this chapter and validate it using a data set containing robot-assisted hysterectomy procedures.

Groundtruth Labeling of Low-level Activity

A common requirement of phase labeling methods is the availability of groundtruth information irrespective of the data modality. Current vision and neural networks based approaches[10,11,135,136,139] for phase recognition are supervised learning methods. They require ground truth annotations of activity segments for learning model parameters, and some even rely on instrument identification information for achieving higher recognition performance. Also, deep learning methods require a lot of training data. Generating groundtruth labels over a large data set and at the scale of surgical procedure videos (that can be two to three hours long) is even more challenging.

On the other hand, approaches using activity tuples of the form (*action*, *actor*, *tool*, *anatomy*),[146–148,150,151] have shown successful validation across multiple surgical procedure data sets and disciplines for phase recognition. Obtaining data on such tuples requires installing multiple sensors in the operating room, and can be prohibitive at scale because of the need for manual annotation during surgery.

In summary, low-level surgical scene and activity context are useful for surgical

phase modeling – whether using the context information as a direct input in formalized and rule-based frameworks or as training input for learning model parameters of vision-based methods to perform automated context extraction. However, gather this context information is challenging.

We believe that crowdsourcing (Section 2.4) can generate surgical context information accurately, reliably and efficiently. We will present pilot validation of this hypothesis.

In the following sections, we will layout the problem statement for surgical phase recognition and define the concept of "summarizations". We will present a pipeline for surgical phase recognition using such summarizations. Then, we will describe two approaches for summarizing surgical procedures using different data modalities. First, we will use readily available sensor data from the dVSS and validate the framework on a hysterectomy procedure data set. Second, we will demonstrate a pilot study to obtain scene context summarizations from a crowd of surgically untrained individuals and perform a pilot validation of phase labeling using such crowdsourced context summarizations on a hysterectomy procedure data set.

We believe valid phase information in complex procedure flows can be obtained from readily available sensor data from the dVSS without any additional equipment installation in the OR. We believe that crowdsourcing of surgical context information can provide reliable data to predict surgical phase labels as well.

3.2 Problem Statement

Let us represent a surgical procedure X using the data available (collected) during the procedure: $\mathbf{D}$ of size $T \times F$, where T are time steps, F is the dimensionality of the data obtained at each time step. Let there be a set of phase classes: $\mathcal{C} = \{1, 2, \ldots, C\}$. Let the constituent phase sequence of the procedure X be $\mathcal{P} = \{p_1, p_2, \ldots, p_M\}$, where M is the number of phases in X. Let each phase p_i be represented as a tuple:

$$\begin{aligned} p_i &= (y_i, t_i, d_i) \quad \text{such that} \\ t_i &= t_{i-1} + d_{i-1} \\ \sum_{i=1}^{M} d_i &= \mathcal{T} \end{aligned} \tag{3.1}$$

where, y_i is the class label of the phase p_i ($y_i \in \mathcal{C}$), t_i is the start time, and d_i is the duration of the phase segment i, and $\mathcal{T}$ is the total duration of the procedure X.

The problem is to find the phase sequence $\mathcal{P}$, given $\mathbf{D}$.

3.3 Framework

Our proposed framework to solve the above problem of surgical phase recognition consists of (1) a summarization module, (2) a phase scoring module, and (3) a phase labeling module.

3.3.1 Summarization Module

We define summarization as a representation of the low-level activity of the surgical procedure. This may include following information:

- actions performed by the surgeon or an assistant,
- purpose of usage of one or more of the surgical instruments,
- surgical objects (needle, retrieval bag),
- environmental changes (smoke, bleeding, irrigation),
- surgeon's hands, instruments and endoscope motion, and
- system usage events (cautery activation, change of instruments).

There are two parameters associated with summarizations – window size and sampling rate. Window size is the duration over which the summarization is obtained and let us denote this using w. Sampling rate is the skip interval at which summarizations are obtained and let us denote this using s. Figure 3.2 shows the two parameters along a timeline for better understanding. w and s can be chosen empirically or through cross-fold validation setups.

As per the notation defined in Section 3.2, we can refer to $\mathbf{D}$ as the summarization data, F as the dimensionality of the summarization and T as the total number of summarization windows over the length of the procedure, such that:

$$\mathcal{T} = s \times T \tag{3.2}$$

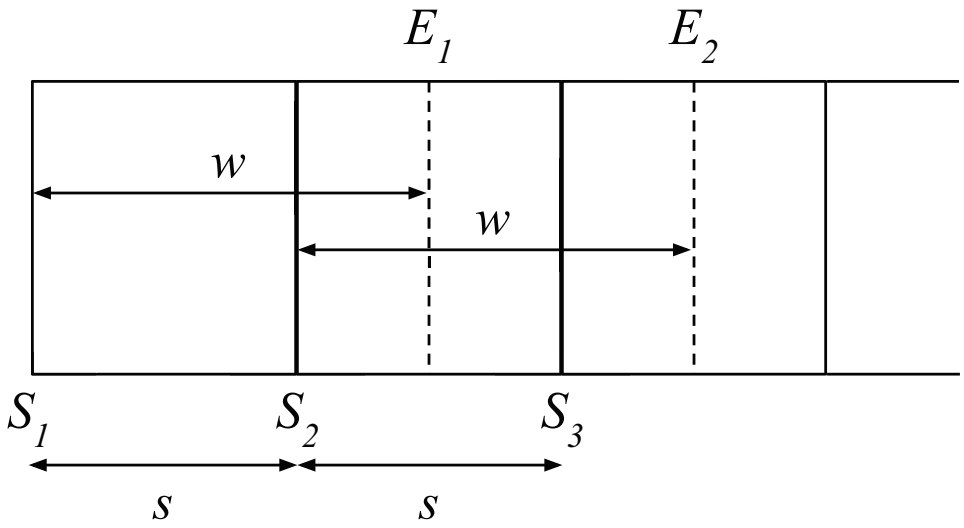


Figure 3.2: Summarization window size and sampling rate example. w is the window size, s is the sampling rate for summarization. $(S_1, E_1), (S_2, E_2)$ are summarization intervals such that $E_i - S_i = w$ and $S_{i+1} - S_i = s$.

3.3.2 Phase Scoring Module

In the above module, we uniformly segment the procedure length into steps at a rate of s seconds (duration) and obtain summarizations of intervals of length w. In this module, a score vector is computed for each row of summarization data matrix ($\mathbf{D}$) which corresponds to the likelihood that the row (time step) D_t belongs to each of the phase classes. Let us denote the score vector as $\mathbf{S}_t \in \mathbb{R}^C$ for the time step t where C is the number of surgical phase classes. Thus, the phase scoring module, PS is as follows:

$$PS(\mathbf{D}_t) = \mathbf{S}_t \qquad \text{such that,} 1 \leq t \leq T \tag{3.3}$$

We compare three scoring models – (1) a linear model applied to features at each time step ($\mathbf{D}_t$), (2) a non-linear model applied to each time step ($\mathbf{D}_t$), and (3) a non-linear model applied to a sequence of time steps ($\mathbf{D}$).

Linear Frame-wise Model

This model assumes there is a linear vector $\mathbf{w}_c \in \mathbb{R}^F$ that discriminates phase c from the rest of the data. Let the score $\mathbf{S}_t^c = \mathbf{w}_c \bullet \mathbf{D}_t$, where $\bullet$ is the dot product operator. If the phase label for the time step $y_t = c$ then the correct score, $\mathbf{S}_t^{y_t}$ should be higher than the score for any other class such that $\mathbf{S}_t^{y_t} > \mathbf{S}_t^c \quad \forall\ c \neq y_t$. In our case, we learn weights $\mathbf{w}$ with a one-versus-all Support Vector Machine (SVM).[152]

Non-linear Frame-wise Model

Each phase may be best classified using a non-linear mapping of the given features in each interval. We follow the work of Stauder et al.[144] who model surgical phase using a Random Forests (RF) classifier. RF is an ensemble learning method that randomly learns which features are most indicative of each class. At each node in the tree, a subset of the features from the training data are randomly selected and tested for their Gini's index as described in the work by Breiman.[153] In our data, we observe different subsets of features are important in characterizing different surgical phase classes. Thus, RF is well suited to our problem. The score for the c^{th} class at time step t is given by the posterior probability $\mathbf{S}_t^c = P(c \,|\, \mathbf{D}_t)$ as computed by this

model.

Non-linear Temporal Model

The previous two models assume that the label at each time step is only a function of the data at the current time step. However, in many phases the features may change substantially between the start and the end of a phase. For example, a surgeon may use monopolar energy at the start of a dissection and a bipolar energy at the end of it. We apply the temporal Convolutional Neural Network (tCNN) of Lea et al.[130] to capture such long-range dependencies across time steps. In the tCNN framework, a set of temporal filters $\mathbf{W}_I \in \mathbf{R}^{h \times F}$ model the features across a sequence of h time steps. Let there be a total of I temporal filters. Each filter models how features change over the course of a phase. The data for each class can be modeled as a function of weights α_i^c that represent how important each filter $\mathbf{W}_i$ is for class c. The score vector is computed as:

$$\mathbf{S}_t^c = \sum_{i=1}^{T} \alpha_i^c \, \mathbf{W}_i * \mathbf{D}_{t:t+h} \tag{3.4}$$

where, $\mathbf{D}_{t:t+h}$ denotes the set of summarizations from time step t to $t+h$. Symbol $*$ refers to a temporal convolution where the summarizations for each time step are convolved over time with the filters $\mathbf{W}$.

3.3.3 Phase Labeling Module

In frame-wise prediction, the class for each time step can be obtained as:

$$y_t = \arg\max_c \mathbf{S}_t^c \tag{3.5}$$

where, y_t is the best scoring phase. While frame-wise accuracy is reasonable, some actions get over-segmented due to high variance in the data in such an approach. We use a segmental inference method based on the semi-Markov Conditional Random Fields to prevent this issue.[154]

Given score matrix $\mathbf{S} \in \mathbb{R}^{T \times C}$, we find the segments $\mathcal{P}$ that maximize the expectation $E(\mathbf{S}, \mathcal{P})$ of the whole sequence:

$$E(\mathbf{S}, \mathcal{P}) = \sum_{i=1}^{M} g(\mathbf{S}, p_i) = \sum_{i=1}^{M} g(\mathbf{S}, y_i, t_i, d_i) \tag{3.6}$$

The segment function $g(\cdot)$ is defined as a sum of the scores within that segment with the constraint that segment i and segment $i+1$ do not belong to the same phase:

$$g(\mathbf{S}, y_i, t_i, d_i) = \begin{cases} \sum_{t=t_i}^{t_i+d_i-1} \mathbf{S}_t^{y_i} & \text{if } y_i \neq y_{i+1}, \\ -\infty & \text{otherwise} \end{cases} \tag{3.7}$$

This model can be viewed in the probabilistic setting as a Conditional Random Field using $Pr(\mathcal{P} \mid \mathbf{S}) \propto \exp(-E(\mathbf{S}, \mathcal{P}))$.

We solve the following discrete constrained optimization problem to find all phase segments, their start time, and their duration:

$$\mathcal{P} = \underset{\mathcal{P}=\{p_1,\ldots,p_M\}}{\arg\max} \; E(\mathbf{S}, \mathcal{P}) \tag{3.8}$$

$$s.t. \qquad t_i = t_{i-1} + d_{i-1}, \qquad \sum_{i=1}^{M} d_i = T \qquad \textbf{and} \qquad 0 < M \leq T$$

In the naive case, this problem has computational complexity $O(T^2C^2)$. We use the method proposed in Lea et al.[130] that is of the order $O(KTC^2)$ where K is an upper bound on the number of segments. K is typically much smaller than T.

A block diagram of our overall framework is shown in Figure 3.3. We have proposed a phase recognition framework using: a summarization module that represents the low level activity of the surgery, a phase scoring module that assigns class scores to each summarization interval, and a phase labeling module that infers the final segment labeling.

3.4 Experiment Setup

In this section, we will describe the data set, implementation of the framework and evaluation metrics used.

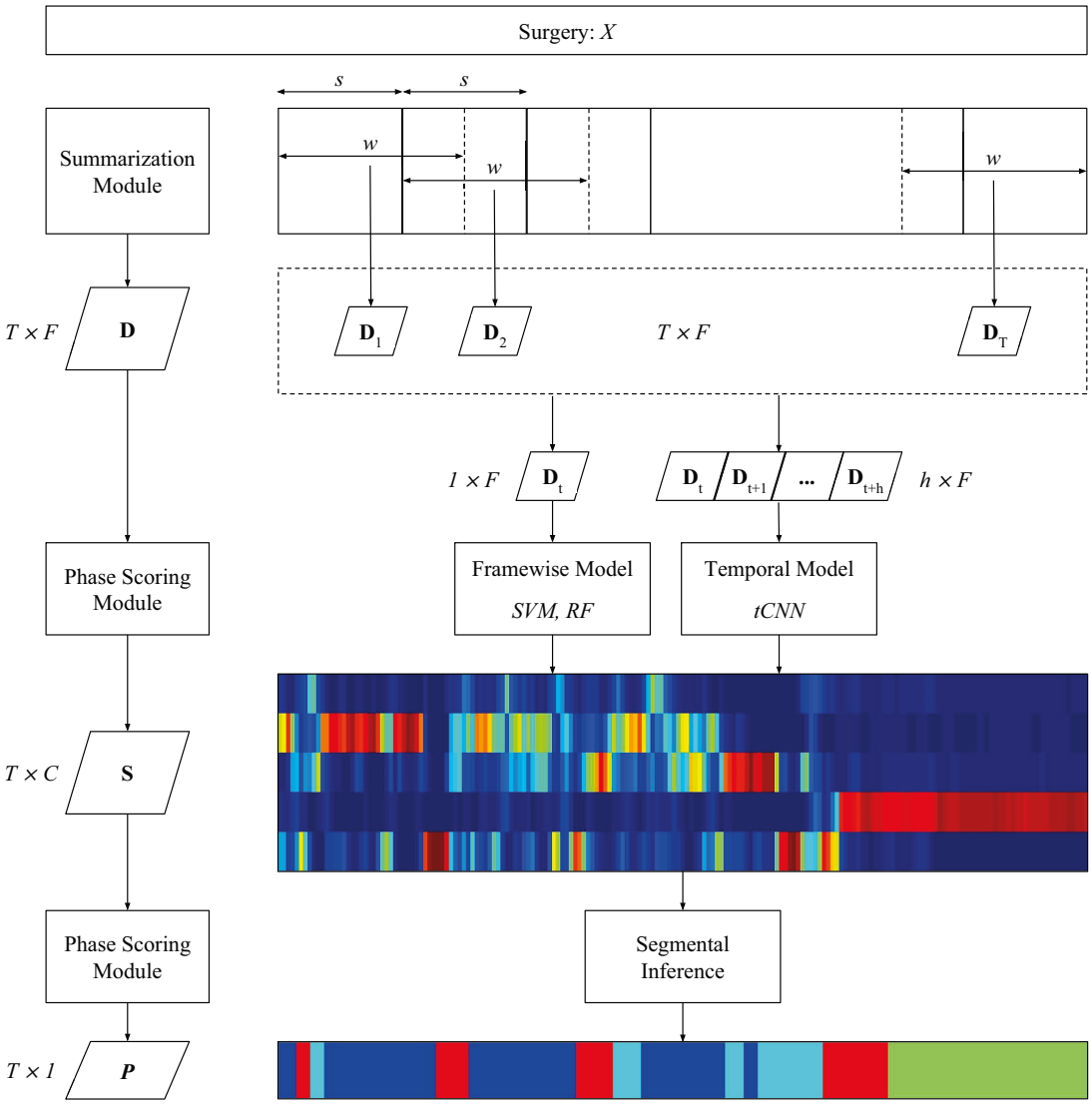


Figure 3.3: Surgical phase recognition framework using summarizations. SVM: support vector machine, RF: random forests, tCNN: temporal convolutional neural network. $\mathbf{D} \in \mathbb{R}^{T \times F}$: summarizations, w: window size, s: sampling rate, T: number of time steps, $\mathbf{S} \in \mathbb{R}^{T \times C}$: phase class scores, $\mathcal{P}$: predicted phase sequence.

3.4.1 Data

We used the WarmUp Hysterectomy data set described in Appendix A.3. We used a subset of 24 RALH (robot-assisted laparoscopic hysterectomy) procedures that were complete. This excludes those recordings that had missing video or system event data.

RALH are highly variable in duration and phase flow. This is unlike procedures like lap-chole which have been studied in many previous phase detection papers. Our data set contains surgeries that range from 47 minutes to 3 hours and 47 minutes in length and contain between 8 and 48 phase instances. Six faculty surgeons performed the procedures with the assistance of 23 surgical residents. At least two surgeons operated in each procedure.

The data set contains activity labels for phases of RALH. The summarizations we plan on using cannot distinguish between anatomical structures, and so similar phases were grouped and merged into higher-level labels as shown in Table 3.5. In total, our system has 5 phase label classes:

$$\mathcal{C} = \{Ligation, Dissection, Colpotomy, Suturing, \text{and } Transition\} \tag{3.9}$$

Skewed Phase Distribution

In RALH, some of the surgical phases are much longer in duration than others. Table 3.5 shows that the ground truth phase distribution ('Prior' column) is highly

Table 3.5: Phases in RALH: after merging original labels.

Original Phase	Merged Phase	Prior
Ligate {IP/Round/UO} {Left/Right}	Ligation	0.07
Isolate Uterus Isolate Ovary Left/Right Dissect Auxiliary Tissue Dissect Lymph Nodes	Dissection	0.36
Colpotomy	Colpotomy	0.07
Suture Vaginal Cuff	Suturing	0.18
Transition Extract Anatomy	Transition	0.32

skewed towards *Dissection* and *Transition* class. To account for this, we sub-sampled the training data for the frame-wise phase scoring models (SVM and RF classifiers) to create a balanced training set. We created 100 iterations for training set in each of the validation folds. The final score $\mathbf{S}_t \in \mathbb{R}^C$ for a test sample was the average of the score over the 100 iterations. However, as the test set was expected to be skewed, the training data class distribution was set as the class weight (priors) for the SVM and RF models.

The most important phase labels from a surgical standpoint – *Ligation* and *Colpotomy* – are sometimes very short. Thus, an analysis on the effect of sampling rate s (Section 3.6) for the summarization on the phase recognition performance should be conducted.

3.4.2 Implementation of Modules

All data was normalized using zero-mean and unit-variance scaling using statistics from the training data. Cross validation was performed to find the hyperparameters in each model. A RF classifier containing 100 trees was chosen based on out-of-bag estimation error over the range of $[10, 500]$ trees. The minimum number of samples at each leaf node was set to 5. The tCNN was implemented using Keras [1], an efficient library for developing deep learning models. We set the filter duration h to be 20 time steps based on cross-validation. For segmental inference, we set the upper bound K on the number of phases in a procedure to be 15.

3.4.3 Evaluation Metrics

We evaluate the framework's performance using an overall accuracy, the per-class precision/recall and a segmental Levenshtein distance. Accuracy, precision and recall are computed using their standard formulae. In case of accuracy, we computed a macro average over each fold of the cross validation setup.

We computed the Levenshtein distance metric (LD)[155] to calculate the different between the groundtruth and predicted phase sequence without accounting for the boundary points (start and end of each phase label). While the accuracy metric penalizes the algorithm for incorrect boundaries, the LD metric penalizes for incorrect ordering of phases. A combination of the two metrics should be considered while

[1] Keras: Deep Learning library: `http://keras.io`

comparing the performance of different prediction models. An algorithm can get high accuracy scores with large number of spurious false positives in its predicted phase sequence compared to an algorithm that does not do so. The LD metric can prove to be useful to do a trade-off between accurate segmentation and accurate ordering.

Levenshtein Distance

This is the difference between two string sequences computed as the minimum number of edits (insertions, deletions and substitutions) that need to be performed to change one sequence into the other. Each set of predictions is split into its constituent segments. For example, "AAABBCCCC" becomes "ABC". The resulting collapsed sequence is compared between the groundtruth and prediction. The LD metric is typically computed in a dynamic programming setting with costs for making each edit. We chose a uniform cost of 1 for insertions, deletions, and substitutions. The number of segments in each prediction and ground truth labeling may vary, thus LD is normalized by the maximum number of segments in each prediction and ground truth labeling and converted to a percentage. Note, smaller values for LD indicate better performance.

3.5 System Events-based Summarizations

In this section, we will present surgical summarization using readily available event data such as a binary signal indicating if an energy instrument is active. We refer to

such data as *system events* or simply *events* in this section. Previous methods (Table 3.4) have used sensor data like carbon dioxide pressure, weight of the irrigation and suction bag, inclination of the surgical table which requires additional, and sometimes sophisticated instrumentation of the OR prior to the surgery. On the contrary, we present our work in the context of the dVSS and rely on the da Vinci API to easily obtain the event data without any additional instrumentation (except for the da Vinci Data Recorder described in Section 2.1.4 which is non disruptive to the OR workflow and does not require any special installation).

3.5.1 Summarization Features

We define a set of features (Table 3.6) that summarize instrument and event information. These features are motivated by the notion that a particular surgical phase must be completed using a specific set of instruments. For example, a suturing phase should ideally be performed using a large needle driver instrument.

We categorize instruments into four types: *monopolar energy instruments*, *bipolar energy instruments*, *needle driver instruments* and *regular (non-energy) forceps instruments*. Note that, while some instruments are intended for cautery actions, there are times when a surgeon will use them for other tasks like grasping or retracting as well.

For cautery tasks, the surgeon uses one form of energy over the other based on the current step and the surrounding anatomy. For example, a surgeon applies bipo-

Table 3.6: System events-based summarization features and their descriptions

Name	Description
	Fraction of segment length for which
MonopolarCutTime	monopolar cut energy was active
MonopolarCoagTime	monopolar coagulation energy was active
BipolarTime	bipolar energy was active
TotalTime	any of the energy types was active
CameraTime	camera was moved
ClutchTime	clutch was pressed
HeadInTime	surgeon was looking into the console
	Number of times
MonopolarCutCount	monopolar cut energy was activated
MonopolarCoagCount	monopolar coagulation energy was activated
BipolarCount	monopolar cut energy was activated
TotalCount	any of the energy types was activated
CameraCount	camera was moved
ClutchCount	clutch was pressed
	Binary flag indicating
IsNeedleDriver	needle driver instrument was in use
IsRegularForceps	non-energy forceps instrument was in use
IsMonopolarTool	monopolar energy instrument was in use
IsBipolarTool	bipolar energy instrument was in use

lar energy to coagulate a structure that is small enough to be grasped between the instrument's grippers. Bipolar energy instruments isolate most of the electrosurgery current passed to the grasped tissue or blood vessel. To contrast, a monopolar instrument is used when dissecting a larger area without any significant anatomical structures or vasculature nearby.

We use other events recorded from the dVSS including instrument installation and removal, camera control, clutching, and head-in indicator (Table 2.2). While evaluating these summarizations, we will present results using the set of events that are common across open, endoscopic and laparoscopic surgery as well as using the entire set of events that are available in the RAMIS setting only.

There are three types of summarization features computed using these events (Table 3.6). The first type are continuous features that are based on duration of an event during the summarization window. Second type are count-based on how many times an event occurred during the window. Third type are binary and based on whether an instrument was in use within the window. Thus, we compute a summarization vector $\mathbf{D}_t \in \mathbb{R}^F$ composed of each item listed in Table 3.6. When using all dVSS events the vector is of length 17. Figure 3.4 shows a temporal plot of the summarization features for a robot-assisted hysterectomy procedure from our data set.

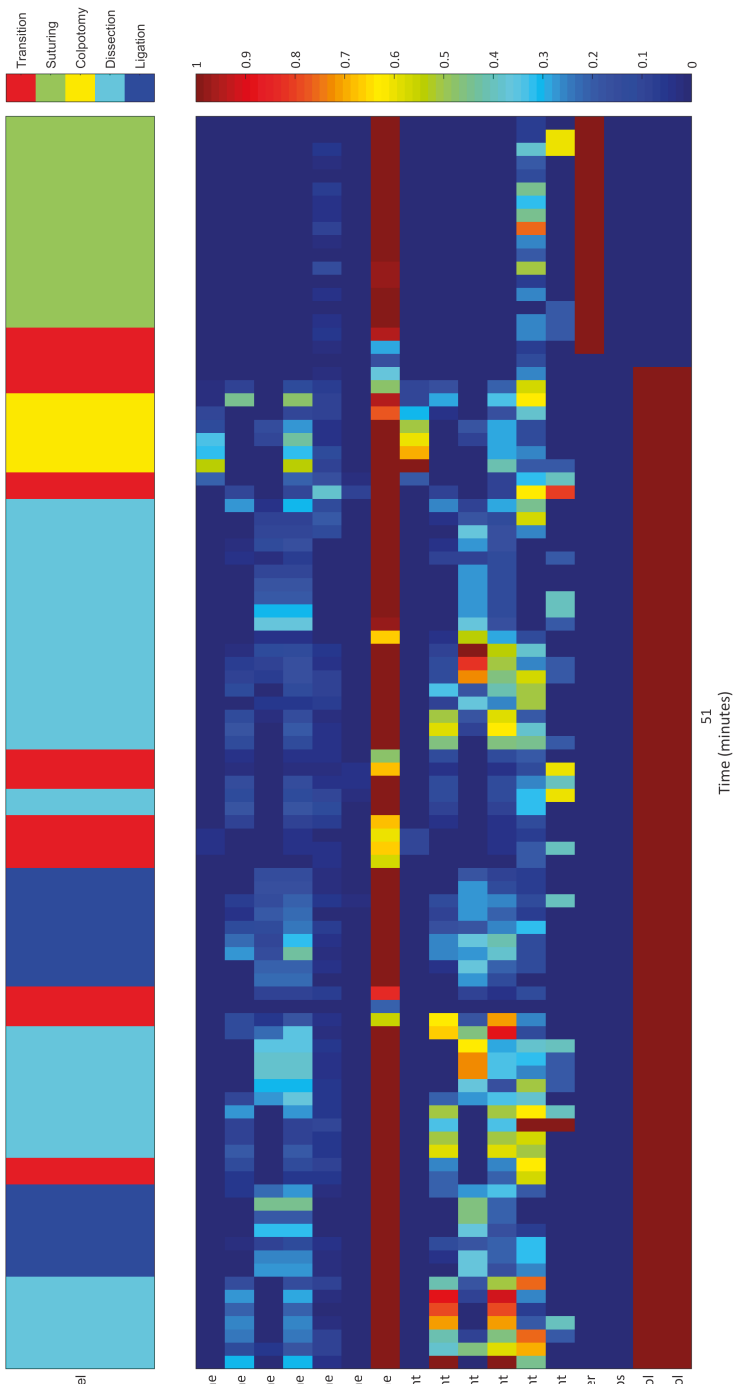


Figure 3.4: System events-based summarization features for a sample hysterectomy procedure from our data set. (Note: feature values have been scaled to [0,1] for better contrast)

3.5.2 Experiments

We use all the 24 RALH procedures in the validation of the system events-based summarizations for phase recognition.

Feature Extraction

In total, the 24 RAH procedures contain approximately 50 hours of data. We choose a summarization window size (w) of length 90 seconds with a sampling rate (s) of 30 seconds resulting in 5781 windows (time steps) across all surgeries. We will show later in the results section, sensitivity analysis on window size (60 to 180 seconds) and sampling rate (10 to 60 seconds). Note, it is possible for a single window to contain more than one distinct phase label. In such a case, the label that spans larger portion of the interval is chosen as that interval's groundtruth phase label.

Sensitivity Analyses

In addition to the validation of the three models, we perform three sets of experiments to analyze the effect on some of the hyperparameters of our framework on the phase recognition performance.

I. Summarization Window Size: This is the time interval over which the summarizations are obtained (Section 3.6). For a window size of 120 seconds, if the bipolar energy tool was activated 10 times during the period $(t, t+120)$ then its count feature at time t would be 10. We evaluate framework performance for window sizes ranging

from 60 seconds to 180 seconds in increments of 30 seconds.

II. Summarization Sampling Rate: This is the time step at which summarizations are computed (Section 3.6). When using a sampling rate of 60 seconds, most instances of the important phases like ligation are contained within a single time step. We evaluate the framework performance at different sampling rates – 10, 30, 45 and 60 seconds.

III. Feature Set: Although our data was recorded using a dVSS, a subset of the features, like those derived from energy activations and instrument identification, can be captured easily and at a low cost using button sensors and RFID tags. These signals are generic across laparoscopic, endoscopic and open surgical procedures. We evaluate our framework's prediction performance using a 12-dimensional subset vector ($EtECtTi$) containing 4 time-based energy features (Et), 4 count-based energy features (ECt), and 4 instrument information flags (Ti).

3.5.3 Results

We compute the framework performance using a leave-one-surgery-out cross validation over all 24 procedures. We address several questions: (1) what is the overall accuracy and precision/recall for each surgical phase, (2) what is the impact of segmental inference, (3) how do the summarization window size and sampling rate impact accuracy, and (4) do signals specific to the dVSS enhance performance versus signals available and generic to most other forms of surgery.

Table 3.7 shows the overall phase recognition performance using frame-wise prediction accuracy and the LD metric (macro averages). Results using frame-wise inference are listed on top and using segmental inference are on bottom. In general, all the three methods perform equally well when comparing accuracy. tCNN does a better job at the sequence prediction compared to RF and SVM (LD metric is lower) when using frame-wise inference. However, these differences disappear when segmental inference was used on top of the phase scores. Segmental predictions more accurate and correctly ordered than the corresponding frame-wise predictions (but by a small margin of 3%). Note, all the results have overlapping standard deviation intervals ($\pm 3\sigma$).

The phase label predictions from the three approaches along with the groundtruth phase sequence from one of the data set procedures is shown in Figure 3.5. Additionally, the feature importance metric computed based on mean-squared error at each node from RF showed that all features were similar in importance.

Tables 3.8 and 3.9 show per-phase precision and recall respectively. Precision is higher for *Suturing*, moderate for *Dissection* and *Transition* and low for *Ligation*. Highest precision values for each class came from a segmental inference prediction. We do not observe any other trend. *Suturing* phase has perfect recall and *Dissection* has recall of around 80%. Recall for *Ligation* was poor in most cases, especially in segmental inference based predictions.

Table 3.7: Phase recognition accuracy for various sampling rate sizes

Method	Accuracy (σ)	LD (σ)
SVM	69.4 ± 7.4	58.9 ± 10.1
RF	71.7 ± 7.6	58.3 ± 11.4
tCNN	74.6 ± 7.5	43.3 ± 14.5
SVM (*Seg*)	72.6 ± 8.5	43.2 ± 12.4
RF (*Seg*)	72.8 ± 9.0	**42.3 ± 13.6**
tCNN (*Seg*)	**75.7 ± 8.5**	42.6 ± 12.3

Summarizations parameters: $w = 90$ seconds, $s = 30$ seconds.
Bold font indicates scoring method with highest accuracy.
(*Seg*) refers to segmental inference-based results.
Smaller values for LD indicate better performance.

Table 3.8: Phase prediction precision per-class

Phase	SVM	RF	tCNN	SVM (*Seg*)	RF (*Seg*)	tCNN (*Seg*)
Ligation	31.7	36.6	49.8	25.8	35.6	**56.5**
Dissection	**70.5**	67.9	69.2	**70.5**	66.5	70.0
Colpotomy	51.0	58.6	61.3	57.3	**71.1**	60.8
Suturing	85.0	84.4	89.9	85.3	82.7	**90.3**
Transition	71.8	77.7	77.4	74.1	77.8	**78.2**

Summarizations parameters: $w = 90$ seconds, $s = 30$ seconds.
Bold font refers to highest precision values across methods
(*Seg*) refers to segmental inference-based results
* segmental inference lowered the precision value

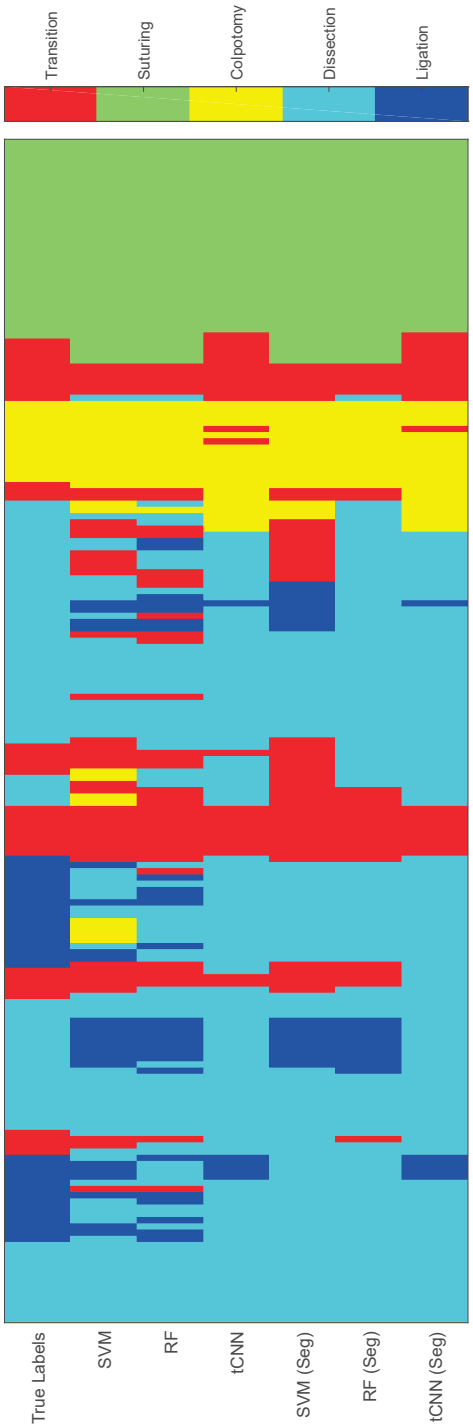


Figure 3.5: Phase recognition for a hysterectomy procedure from our data set using system events-based summarizations. (*Seg*) refers to segmental inference-based predictions

Table 3.9: Phase prediction recall per-class

Phase	SVM	RF	tCNN	SVM (*Seg*)	RF (*Seg*)	tCNN (*Seg*)
Ligation	34.8	**37.0**	35.4	13.3	18.8	31.2
Dissection	57.7	68.8	77.8	70.5	79.8	**82.1**
Colpotomy	59.2	54.8	58.1	**62.2**	48.4	53.7
Suturing	97.2	99.6	94.5	98.6	**100.0**	96.4
Transition	**76.4**	69.1	70.2	74.3	64.6	69.9

Summarizations parameters: $w = 90$ seconds, $s = 30$ seconds.
Bold font refers to highest recall values across methods
(*Seg*) refers to segmental inference-based results

Sensitivity Analyses

Table 3.10 shows effect on accuracy in phase prediction as part of the first sensitivity analysis (Section 3.5.2) using features computed with summarization window size varying from 60 to 180 seconds. The performance is similar among all values, however, results at 60 seconds are marginally worse. This matches our intuition to choose a window size of 90 seconds for the main results based on the typical phase lengths for hysterectomy procedures.

Table 3.11 shows effect on accuracy in phase recognition for different summarization sampling rates. It shows that there is a minor increase in accuracy as the sampling rate decreases from 60 to 10 seconds. The results stabilize around 30 seconds. This may be because phases with short duration, such as *Ligation*, yield a small number of samples.

Table 3.12 compares results using all signals (events) recorded from the dVSS ver-

Table 3.10: Phase recognition accuracy using different window sizes for summarization

Method	Window Size (seconds)				
	60	90	120	150	180
SVM (*Seg*)	**73.1 ± 8.8**	72.6 ± 8.5	71.2 ± 8.5	70.7 ± 10.1	71.6 ± 9.4
RF (*Seg*)	74.3 ± 8.2	72.8 ± 9.0	73.4 ± 8.3	**74.3 ± 7.8**	73.9 ± 7.4
tCNN (*Seg*)	**76.5 ± 8.3**	75.7 ± 8.5	74.5 ± 8.8	75.2 ± 8.7	73.7 ± 7.5

Sampling rate was fixed at 30 seconds.
Bold font indicates window size with highest accuracy for each method.
(*Seg*) refers to segmental inference-based results.

Table 3.11: Phase recognition accuracy using different sampling rates for summarization

Method	Sampling Rate (seconds)			
	10	30	45	60
SVM (*Seg*)	72.2 ± 8.7	**72.6 ± 8.5**	69.8 ± 9.2	69.5 ± 9.0
RF (*Seg*)	**75.0 ± 6.9**	72.8 ± 9.0	72.1 ± 9.9	71.8 ± 8.7
tCNN (*Seg*)	75.3 ± 8.5	**75.7 ± 8.5**	74.8 ± 9.6	74.0 ± 7.6

Window size was fixed at 90 seconds.
Bold font indicates sampling rate with highest accuracy for each method.
(*Seg*) refers to segmental inference-based results.

Table 3.12: Phase recognition accuracy using signals specific to the dVSS (*all*) versus signals generic to most types of surgeries (*EtECtTi*).

Feature Set	SVM (*Seg*)	RF (*Seg*)	tCNN (*Seg*)
all ($\mathbb{R}^{17}$)	72.6 ± 8.5	72.8 ± 9.0	75.7 ± 8.5
EtECtTi ($\mathbb{R}^{12}$)	66.3 ± 9.3	70.5 ± 9.0	74.3 ± 8.3

Summarizations parameters: $w = 90$ seconds, $s = 30$ seconds.
EtECtTi: refer to Section 3.5.2
(*Seg*) refers to segmental inference based phase predictions

sus the subset features *EtECtTi* that are common to most types of surgeries (Section 3.5.2). Our results show that performance using these generic features is only a small amount worse than using all features.

3.5.4 Discussion

Earlier, we mentioned that current methods for surgical phase recognition have focused on laparoscopic cholecystectomy procedures which are sequential in nature. Our hysterectomy data set is representative of complex procedures and contains natural variations in procedure flow pertaining to patient anatomy, type of hysterectomy (total, radical, subtotal) and surgeon style. Despite these challenges, the performance of our framework was comparable to the overall accuracy of other reported results using sensor data.[144,145] Precision and recall values across phases have a similar pattern to those reported by Stauder et al.[144] – dominant phase classes tend to have a much higher precision and recall than other classes.

Despite investigating several models with various distinct assumptions, we found

all approaches achieved relatively similar performance (in the range of 70% to 75% overall phase prediction accuracy). The first (SVM) assumed a simple linear model, the second (RF) learned the most important subsets of features for each phase, and the third (tCNN) non-linearly modeled the temporal evolution of features. We believe that such performance is not a result or limitation of the activity models chosen, but inherent to the data set's temporal variability in terms of the features chosen. While we hypothesized and showed that system events-based summarizations containing information about energy (cautery) usage and instrument identification are useful in predicting the surgical phase, the different phase classes in this feature space show similar temporal variations. For example, the dissection phase contains coagulation and cutting of the uterine blood vessels which has a very similar pattern to the phase of ligation wherein the ligaments are coagulated and cut. As a result, the short duration phase instances of ligation merge into the nearby larger instances of dissection during the segmental inference of phase labels. Similarly, the tCNN's temporal filters smooth out feature responses across these smaller phases leading to lower precision and recall values.

While our primary validation was based on events data captured from a robot-assisted surgery platform, we performed the same experiments by leaving out some of the robot-specific events like camera control, clutching, and the console head-in sensor. This analysis showed that performance of the different models in predicting phase labels did not decrease significantly using the smaller set of features generic to

other forms of surgery (Table 3.12). Thus, our method can be applied and tested with non-robotic surgical systems. Previous works[140,144] have successfully captured these signals in the lap-chole procedure settings. This would enable large scale studies that require surgical phase analysis in the domain of traditional laparoscopic as well as open surgery, in addition to robot-assisted procedures.

System events are one of many data modes to represent summarizations. Next, we will venture into crowdsourcing and surgical context based summarizations with the same end-goal of surgical phase recognition.

3.6 Crowdsourced Surgical Context Summarizations

In this section, we will present summarizations based on surgical context information. We will use crowdsourcing to obtain these summarizations with a dual objective of performing phase recognition and generating reliable and valid training data sets for future video-based activity modeling methods.

Previous applications of crowdsourcing in the biomedical domain have shown success in understanding protein structures,[17] diagnosing disease,[21,156] evaluating surgical skill,[23] annotating instrument boundaries in laparoscopic surgery images[28] and identifying polyp-free segments in colonoscopy videos.[26]

3.6.1 Data Pre-processing

We used the *WarmUp Hysterectomy* data set (Appendix A.3) for this study. We chose the subset of 29 procedures that contained video, motion and system events data. A typical RALH procedure lasts about 90 minutes and our goal was to obtain low-level context summarizations.

We surveyed a group of our laboratory members to decide the window size for summarizations to be obtained from the crowd. We showed 5 individuals three video segments of 60, 90 and 120 seconds duration. These segments were selected at random from a surgery video. We asked each of them to summarize the videos independently as well as to indicate what duration of video they preferred to perform such a summarization. Based on this feedback, we chose to provide the crowd with 90-second segments from the procedure videos, and at a sampling rate of 60 seconds (thereby, having a 30 second overlap between two temporally adjacent video segments). We obtained 2914 segments from the 29 procedures using this sampling rate.

Since the WarmUp Hysterectomy data set is still under an active IRB study we went through an approval process to use the video data for crowdsourcing. We anonymized the video images by re-writing them to replace surgeon identifiers with black pixels, and by removing all portions of the video showing images outside of the patient's body. `FFmpeg`[2] (a command line utility for video processing) was used to perform this anonymization. The video frames wherein the endoscope was out of the

[2]https://ffmpeg.org/

patient's body were replaced with black frames containing the centered text "`CAMERA OUT OF PATIENT BODY`".

3.6.2 Crowdsourcing Protocol

We used MTurk (refer to Section 2.4 for explanation of the crowdsourcing terms) for crowdsourcing. In addition to the regular worker qualifications (Section C.3), we created a custom qualification (CQ) to train and orient the workers for the context summarization task. We required workers to pass a test HIT on material covered during training to obtain this CQ. Workers with this CQ were eligible to participate in the final survey HIT. Figure C.2 shows the protocol snapshot as explained in the following paragraphs.

Qualification Training

During the training, the worker was asked to watch a 5-minute video explaining the following concepts:

- RALH procedure overview (describing what part of the body and what organs are seen in the videos)
- activity segments – sharp dissection, blunt dissection, coagulation, suturing, anatomy extraction, exploration, transition and idle (explaining the characteristic action related to them)

- robotic and non-robotic (laparoscopic) instruments (describing the difference between the two and their respective use case during the surgery),
- surgical objects like needle, gauze, retrieval bag, cervical ring (showing images for each of them and describing what they are used for)

The workers were required to watch the video at least once but they were free to watch it as many times as they wished before taking the test. The training video also included an attention phrase, which the workers were required to observe, remember, and enter during the test. Figure 3.6 shows a snapshot of the training page. In addition to the training video, details about the user study and the informed consent were displayed on the landing page as per IRB requirements. The information from the training video was available under the different tabs on top of the page for quick referencing.

Qualification Testing

Upon watching the training video completely, a link to the test page was shown to the worker. The qualification test included the following questions:

1. Please enter the attention phrase from the training video.
2. Select all the images that contain a robotic instrument (four thumbnails of robotic and non-robotic instruments were shown).
3. What is the activity being performed during the video shown on the left?

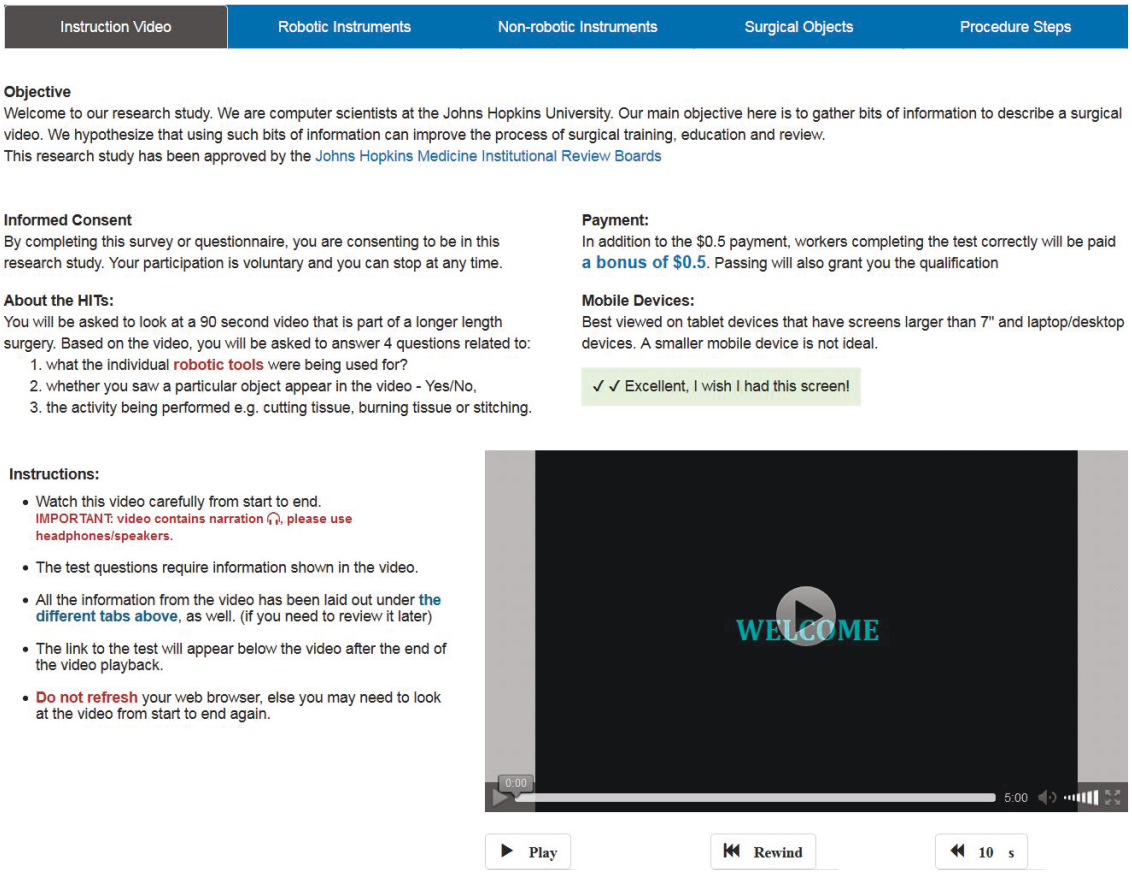


Figure 3.6: A snapshot of the custom qualification training page.

4. What are the robotic instruments being used for in the video shown on the left?

The training page link was available to the workers on the test page for reviewing any concepts while answering the test questions. We gave three attempts to each worker to pass the test, since we realized that by doing so the worker will learn about the task better compared to directly failing them. The workers were not aware of the number of attempts they had. If the workers' answers to any of the questions were incorrect, they were informed of the question they had failed upon, as a reinforcement to understand where they went wrong. Upon a successful completion of the test, the worker was granted the CQ with a score of 1.

The entire qualification training and test framework was part of a HIT with a reward of \$0.5. Workers who passed the test received a bonus of \$0.5 for doing so.

Summarization HIT

A snapshot showing the layout of the HIT is shown in the Figure 3.7. The video to be summarized is shown on the left[3], survey questions on the right. Navigation tabs and buttons were provided to jump through the different summarization questions. Workers were required to watch the complete video before they could submit their responses. Although, they could check and mark their responses while watching the video.

The following questions were asked about the surgical context (phrases in the paren-

[3]The layout was responsive and changed based on the device's screen size.

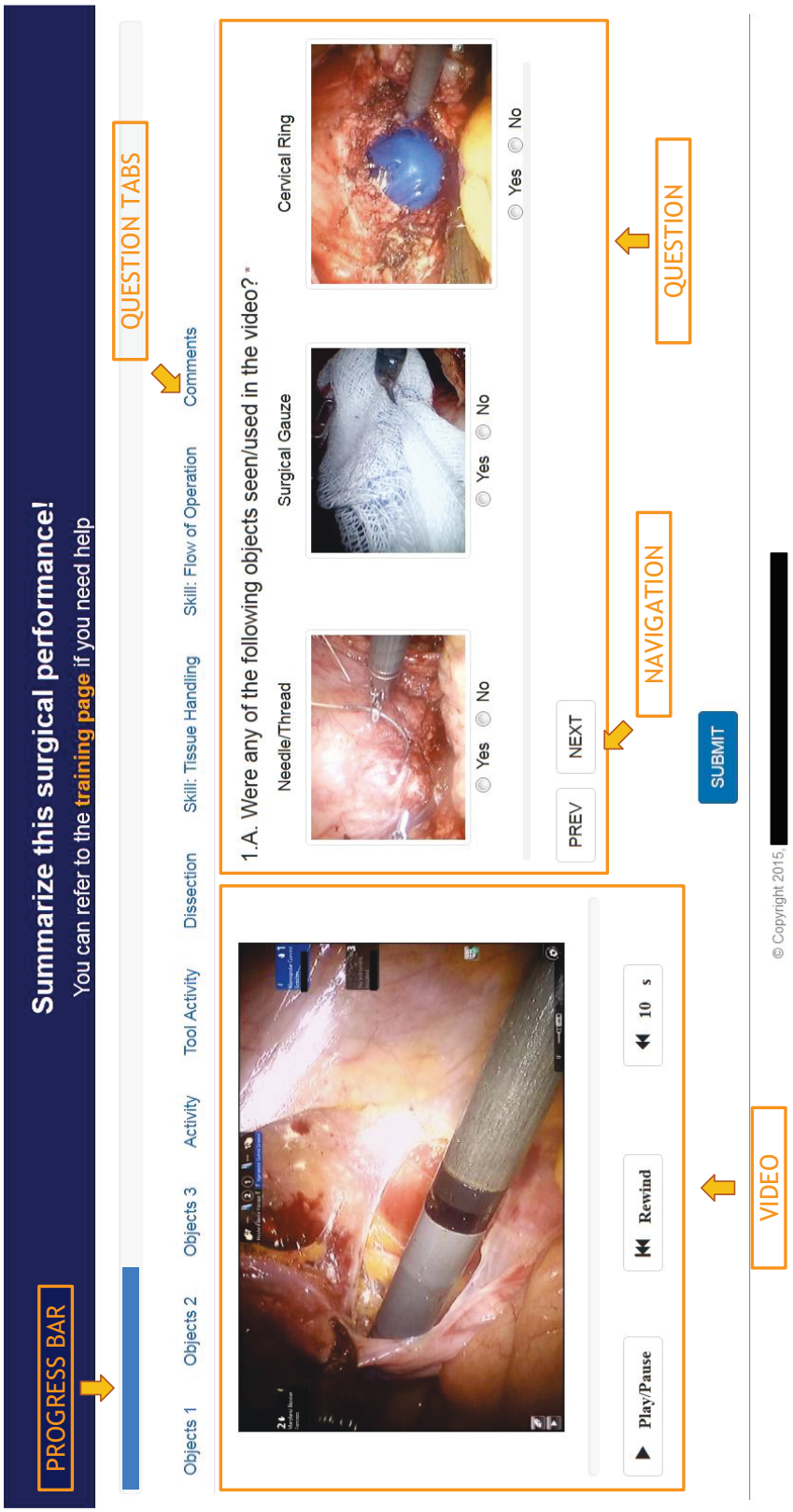


Figure 3.7: A snapshot of the context summarization HIT

theses were the answer options):

1. Were any of the following objects seen/used in the video? (needle, surgical gauze, cervical ring, suction tool, plastic bag, laparoscopic grasper)
2. Did you see any of the following events happen in the video? (bleeding, water, excessive smoke)
3. Which of the following activity is being performed? (sharp dissection, blunt dissection, coagulation, suturing, anatomical extraction, exploration, transition, idle)
4. What are the robotic instruments being used for? (idle / not visible, holding object / tissue / suture, cut tissue, touch+burn tissue, grasp+burn tissue)
5. If the main activity was dissection, was dissection performed in one or multiple regions? (one, multiple, n/a)
6. Comments (if any on the video)

Additionally, we inserted attention videos into the HITs to discriminate between genuine responses and spam. On such videos, the worker's response was verified against a stored answer key. Upon failing the attention test, their HIT assignment was ended with a message about the failure and their CQ was revoked (Figure C.2).

3.6.3 Pilot Studies

Crowdsourcing of surgical context summarization is a novel approach with little prior data to inform study design. Thus, we conducted two pilot studies:

1. to estimate the optimal number of workers (N) to generate a low-level summary of surgical procedures (*Pilot I*), and
2. to assess whether such a summary contains information not captured in other modes of data, i.e., system events (*Pilot II*)

All the summarization HITs were set with a reward of $2.25 and a duration of 60 minutes along with the default qualification requirements (specified in Section 2.4.

Pilot I

We launched a pilot study with 40 video segments (90 seconds in length) that were sampled at random from the available 2914 video segments. We obtained 25 summarization responses on each of the 40 videos. The users could perform as many different HITs as were available. It was made sure that each worker responded only once per video. Additionally, we collected summarization responses on the 40 videos from one of the study coordinators.

Reliability

We measured the inter-rater reliability using percent agreement as defined in Section 2.5 separately for each summarization question. We repeatedly (100 iterations)

sub-sampled (with replacement) responses for different values of N (number of workers) and computed the mean and variance of percent agreement in their responses for each question in the HIT.

Validity

We pooled the crowd responses using a simple majority and compared to the responses obtained from the study coordinator. We computed percent agreement (Section 2.5) for all the questions.

Pilot II

We hypothesized that crowdsourced context summarization contains information about the surgical scene that is not part of other modes of data such as system events. Thus, we assessed whether such context summarization yielded better classification accuracy compared with system events captured during RALH. We used the system events summarizations described in Section 3.5.1. We selected four procedures (that were between 1 and 1.5 hours in duration) from our data set for this purpose. A total of 285 video segments (90 seconds in length) were obtained from these 4 procedure videos. We used a random forest (RF) classifier with similar parameters as in Section 3.4.2. We computed the phase classification accuracy with a leave-one-procedure-out cross validation setup. We compared three sets of features – context summarizations, system events summarizations and combined context and system events summariza-

tions.

On MTurk, we adopted a different strategy to reduce the total amount spent. We created 50 HITs in all - 15 HITs containing 5 videos (out of the 285) + 1 attention video, and 35 HITs containing 6 videos from the set of 285. Although these 35 HITs didn't contain any attention questions, we wanted to make sure the responses were genuine or from trusted workers. So, we added another qualification requirement to these 35 HITs – CQ score must be greater than or equal to 5. And, we announced to the workers that with every two HITs completed successfully, their CQ score will increase by 1. By doing so, we hoped that workers with more experience and lower likelihood to spam will attempt the 35 HITs without requiring the attention checks.

3.6.4 Results

In the qualification stage of the crowdsourcing, we observed that the percentage of workers who passed the test increased from 6% to 50% after giving them three attempts to answer the test questions correctly.

Figure 3.8 illustrates the crowdsourced surgical context summarization for one of the RALH procedures. The summarization was obtained by majority pooling of crowd responses for each 90 second window. These summarizations were stacked to create the temporal layout for the surgery. An immediate observation is the correspondence between the presence of a needle and the suturing phase. Such context is absent in motion and system events and hard to extract from videos. For objects,

Table 3.13: Percent Agreement in crowd responses per question ($N = 25$).

Question	Percent Agreement
Activity	0.89 (0.10)
Left Instrument Activity	0.95 (0.07)
Right Instrument Activity	0.88 (0.14)
Third Instrument Activity	0.85 (0.15)
# Dissection Regions	0.74 (0.18)
Needle	1.00 (0.01)
Suction Tool	0.97 (0.07)
Laparoscopic Grasper	0.73 (0.14)
Cervical Ring	0.98 (0.07)
Plastic Bag	1.00 (0.00)
Surgical Gauze	1.00 (0.01)
Excessive Smoke	0.90 (0.15)
Water	0.99 (0.04)
Bleeding	0.91 (0.12)

Numbers in parentheses are standard errors.

blue indicates absent in the scene. The activity color coding is: sharp dissection (blue), coagulation (cyan), suturing(green), anatomical extraction (yellow), exploration (orange), transition (bright red), idle (red). The instrument activity is: idle (dark blue), holding object or tissue or suture (blue), cut tissue (cyan), touch+burn tissue (green), and grasp+burn tissue (yellow).

Pilot I

We observed high inter-rater agreement in the crowd's responses to the different summarization questions as shown in Table 3.13.

Table 3.14 shows the estimated percent agreement for a range of values of N (num-

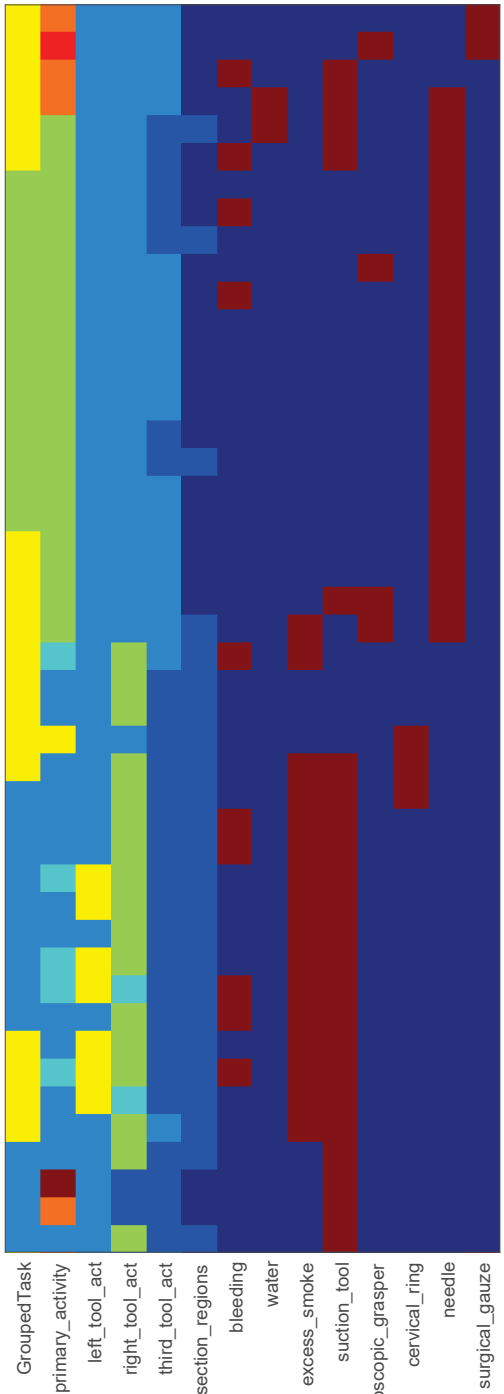


Figure 3.8: Crowdsourced surgical context summarization for a RALH procedure. The groundtruth surgical phase is shown in the top row. Color legend is described in Section 3.6.4

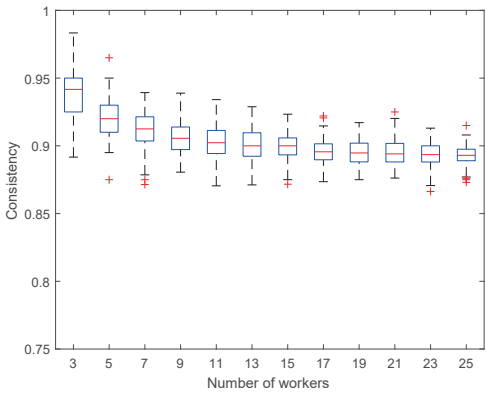
Table 3.14: Percent agreement in responses versus N (numbers of workers)

	Number of Workers			
Question	5	10	15	20
Activity in the video	0.92	0.91	0.90	0.90
Left Instrument Activity	0.96	0.96	0.96	0.96
Right Instrument Activity	0.91	0.90	0.90	0.90
Third Instrument Activity	0.48	0.43	0.41	0.41
# Dissection Regions	0.80	0.77	0.77	0.76
Suction Tool	0.97	0.97	0.97	0.97
Laparoscopic Grasper	0.78	0.75	0.74	0.74
Cervical Ring	0.99	0.99	0.98	0.98
Excessive Smoke	0.92	0.91	0.91	0.91
Water	0.99	0.99	0.99	0.99
Bleeding	0.92	0.91	0.91	0.91

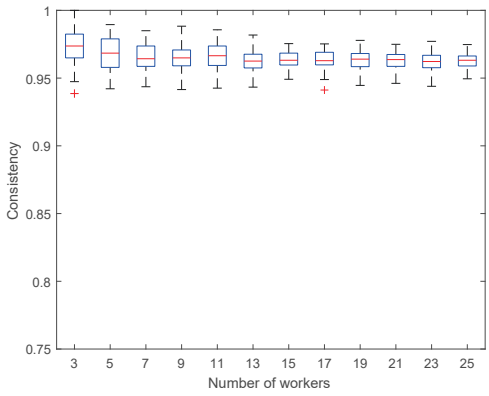
Percent agreement for needle, surgical gauze and plastic bag were at 1.0 throughout the range of N.

ber of workers) by subsampling from the pool of 25 responses. Percent agreement estimates using a finer increment in N are shown in Figure 3.9. Our findings suggest that 5 workers can provide summaries of videos with high consistency on most questions related to activity, instrument usage and objects.

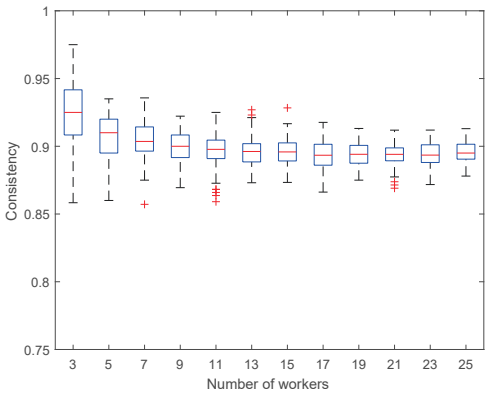
The validity of the crowd's responses compared to reference responses provided by one of the study coordinators are shown in Table 3.15. Pooled crowd responses had a high validity for object and events questions 0.93 to 1.00. The validity on activity in the video was 0.73, while the agreement on individual instrument activities were 0.83 and 0.8 (left and right respectively). The agreement on the number of dissection regions was 0.7.



(a) Activity



(b) Left tool activity



(c) Right tool activity

Figure 3.9: Percent agreement estimate versus N (number of workers)

Table 3.15: Validity of crowdsourced context summarizations.

Question	Validity
Activity	0.73
Left Instrument Activity	0.80
Right Instrument Activity	0.80
Third Instrument Activity	0.90
# Dissection Regions	0.67
Needle	1.00
Surgical Gauze	1.00
Cervical Ring	0.97
Suction Tool	1.00
Plastic Bag	1.00
Laparoscopic Grasper	0.93
Bleeding	0.93
Water	1.00
Excess Smoke	0.97

Pilot II

Phase classification accuracy in RALH was 0.65 using context summarizations, 0.45 using system events-based summarizations, and 0.63 using a combined set of summarizations.

3.6.5 Future Work

We demonstrated that a crowd of surgically untrained workers can yield a highly reliable and valid low-level summarizations of context in surgical procedures. We

observed only a moderate accuracy (65%) in classifying surgical phases in RALH. However, this may be due to the small subset of four procedures that were used in Pilot II study. We can say so, because the system events-based features performed quite low (45%) as well compared to the results seen in Section 3.5.3. A future validation using the entire data set is needed to verify the above observations. The success of context summarization has been shown on hysterectomy procedures. Future studies extending this framework to other procedures should also be undertaken to confirm the reliability observed in the above experiment.

3.7 Summary

In this chapter, we described different aspects of surgical activity modeling *viz.* significance and need for automation, activity information in different data modalities, existing works in bench-top simulation and OR procedures. We focused on the problem of surgical phase recognition in procedures. We have presented a scalable solution for phase recognition using summarizations that can be obtained from intra-operative data.

Previous works on phase recognition have shown validation using data sets containing the straightforward procedure of laparoscopic cholecystectomy. We have shown first results on surgical phase detection using a more complex procedure like hysterectomy. Groundtruth availability limits other approaches for surgical phase

recognition using video data and low-level activity tuples. We have shown preliminary validation of crowdsourcing such groundtruth surgical activity context.

Specifically, we demonstrated that system events-based summarizations contain surgical phase information. Additionally, we described a pilot study on crowdsourcing surgical context and activity summarizations. We showed that a crowd provides reliable and valid summarizations of context in surgical procedures.

Chapter 4

Surgical Skill Assessment

What you cannot measure, you cannot improve.

Lord Kelvin

Work from this chapter has been previously published in a peer-reviewed conference proceeding: A. Malpani, S. S. Vedula, C. C. G. Chen, and G. D. Hager, "Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task" in Information Processing in Computer-Assisted Interventions, Springer International Publishing, 2014, pp. 138147[157] and a journal article: A. Malpani, S. S. Vedula, C. C. G. Chen, and G. D. Hager, "A study of crowdsourced segment-level surgical skill assessment using pairwise rankings" Int J CARS, vol. 10, no. 9, pp. 14351447, Jun. 2015.[115] The study protocol and hypothesis testing was done by S. Vedula and A. Malpani (myself). I was responsible for the web development, data set and database maintenance as well as conducting the user study. C. C. G. Chen was the clinical advisor on the existing skill assessment methods and provided feedback on the different rating approaches for segment-level evaluation. G. D. Hager was the technical advisor and principal investigator of the project. He was instrumental in the co-development of the pairwise comparisons strategy for ranking surgical skill.

Table 4.1: Skill assessment is an integral component of a surgical coaching framework

Training Type	Activity Recognition	Skill Assessment	Feedback	Demonstration
VR	✔	🎓[4,5]	🎓[5]	🎓[5]
Bench Top	✔	🎓[4]	✖	✔
OR	✔ 🎓[3]	✔	✖	✖

✔: indicates previous work in literature exists, 🎓[X]: indicates a solution is presented in Chapter "*X*" of this thesis, ✖: indicates no prior work exists and none is presented in this thesis.

Evaluating and quantifying someone's performance is the key to guiding, mentoring and coaching them to improve on that skill. This ties to the coaching activity listed in Section 1.3 – `EVALUATE` (*how did I do it?*) and `CRITIQUE` (*where was I wrong?*). We talked about the importance of understanding and solving the 'where' part of the question in the previous chapter as we gave examples of activity modeling. But, once you have a way to get to the *where*, you also want to find *how* will you measure (quantify) the performance. In this chapter, we will address this question. Before that, we will provide a background on the current state of skill assessment and existing approaches for manual and automated objective skill assessments.

4.1 Background

Formalized surgical education and training can be traced back to the early decades of the 20th century.[158] But, it was not until the late 1990s that standardized assessment of surgical skills for residents (trainees) became an active area in the surgical

education community.

4.1.1 Checklist-based Methods

Martin et al.[1,159] built an objective and structured technical skills assessment tool – OSATS (Figure 2.14) that consisted of a procedure-specific checklist and a generic global rating scale (GRS). This tool has been validated for skill assessment in a large number of studies. On similar lines, a variety of assessment tools using checklists and GRS measures have been developed since. Of these, GOALS by Vassiliou et al.[52] and GEARS by Goh et al.[2] (Figure 2.15) have been validated the most in their specific domains of laparoscopic and robotic (robot-assisted) surgical skills respectively.

Other than being a GRS and a Likert-like rating scale of 1 to 5, another commonality of the skill components on these assessment tools is that they are applicable at the global (procedure or task) activity level. Additionally, these tools were developed with the goal to provide an objective evaluation unlike previous methods where a single score by the grader would determine skill. Subjectivity and bias from the grader about what aspect of skill to assess was removed by providing components on which skill should be rated e.g. tissue handling, flow of operation, bimanual dexterity. However, this approach still suffers from grader bias arising due to a possible tendency to score everyone low or high and to score based on personal preferences. These tools also require watching a live performance or reviewing a recorded performance which means the equivalent amount of time is needed from the grader (attending surgeons).

This makes the use of such tools highly unreliable and invalid as surgeons either skip through the video to reduce time needed per performance or just do not provide their ratings. At the same time, a surgeon's time is more monetarily favorable for hospital administration if spent on surgery and clinic duty. Because of such limitations, automation of skill assessment has been recommended by surgical educators[77] to provide efficient and objective evaluation of surgeons.

4.1.2 Automated Methods

In the early 90s, computer integrated surgery (CIS) was gaining increased attention and CIS devices were making their way into the OR. With this, access to intraoperative surgical performance data was starting to become easier. Instrumenting training labs to capture video and motion data to analyze performance was considered. Researchers like Datta et al.[53,106] and Dosis et al.[107] performed motion analysis to assess skills. Law et al.[160] looked at eye-gaze tracking as measures of surgical expertise. Rosen et al.[57,108] approached the skill classification problem using hidden Markov models (HMMs) and showed validity in predicting the skill class – novice versus expert. Following this, there have been an array of works using machine learning and statistical modeling approaches for surgical skill assessment and classification using motion data.[56,59–61,109–111,161–167] Methods using other modes of data for skill assessment have been developed as well – eye-gaze tracking,[161,162,168,169] instrument vibrations,[170] and videos.[63,64] Additionally, other works[54,55,105,171,172] have looked at

motion metrics along with machine learning and statistical modeling to evaluate surgical skill. Like the checklist-based GRS methods, all the above automated solutions predicted skill class or skill score at the global (task) level.

While automation and objectiveness are desirable, just a skill score or expertise label is not sufficient to train or coach surgeons. The question of *where was I wrong?* still remains unsolved.

4.1.3 Segment-level Evaluation

We should mention that Reiley et al.[110] did look at skill models for gesture level of activity using HMMs. They trained three HMMs per gesture to model novice, intermediate and expert behaviors respectively. Pre-segmented gestures from a new performance were assigned one of the skill classes based on the likelihood under the three models. An overall task-level skill label was determined by taking majority over the gesture-level predictions. Though, an underlying assumption of this work does not always hold true. The expert-level gesture HMM was trained using performance data obtained from a gesture belonging to an overall task-level expert performance. Similar approach was adopted for learning the intermediate and novice HMMs. However, it is not necessarily true that a task-level expert performance is expert-like at all the constituent gesture performances. It may have worked out to be so in the experiments conducted by Reiley et al. But, we see a contradicting scenario in Figures 4.1, 4.2 and 4.3, wherein similar maneuver-level metric values (time, path length

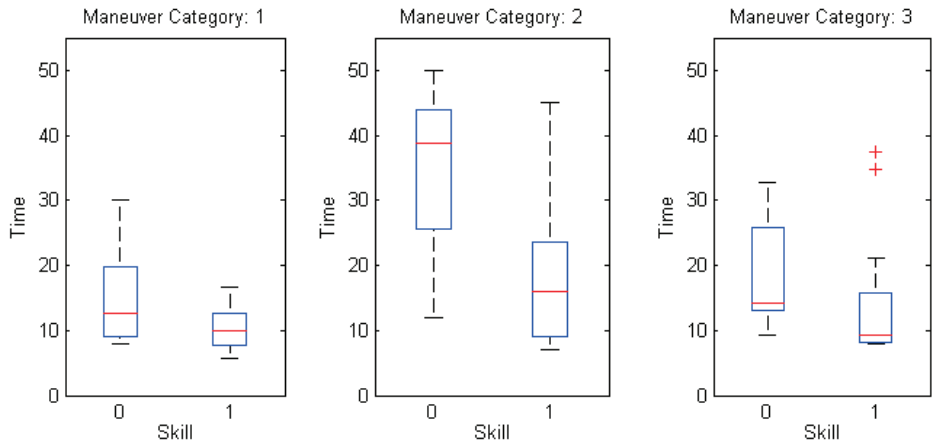


Figure 4.1: Time to complete maneuvers in an interrupted suturing task. Task performances with low OSATS score (≤ 10) are marked as '0' on the x-axis, while those with a high OSATS score (≥ 27) are marked as '1'. Note: OSATS $\in [5, 30]$. © Springer International Publishing Switzerland 2014

and number of movements) are observed for performances of an interrupted suturing task with extreme OSATS-based task-level scores. These figures also motivate the need for segment-level evaluation since task-level assessments (OSATS) alone can be misleading. Vedula et al.[105] show that task-level OSATS-based scores predicted using motion analysis at task-, maneuver- and gesture-levels are equivalent. This additionally supports the notion of segment-level evaluation.

Because segment-level evaluation will not only be useful to provide targeted feedback about which segments need improvement but can lead to reliable task-level evaluations as well.

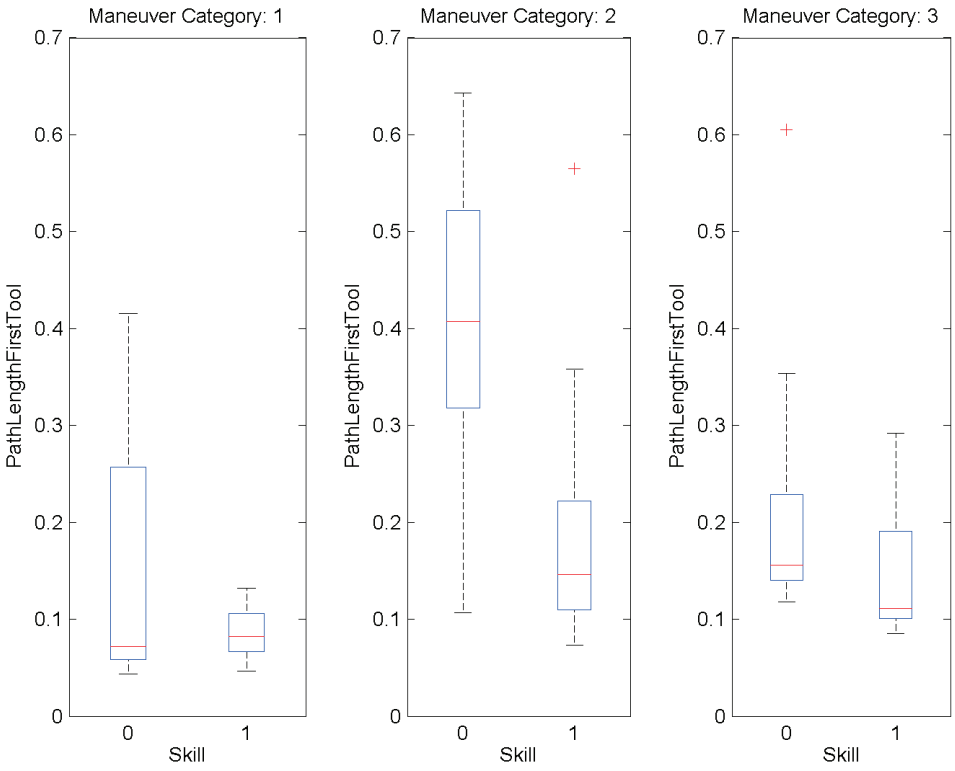


Figure 4.2: Path length traveled by instrument to complete maneuvers in an interrupted suturing task. Task performances with low OSATS score (≤ 10) are marked as '0' on the x-axis, while those with a high OSATS score (≥ 27) are marked as '1'. Note: OSATS $\in [5, 30]$.

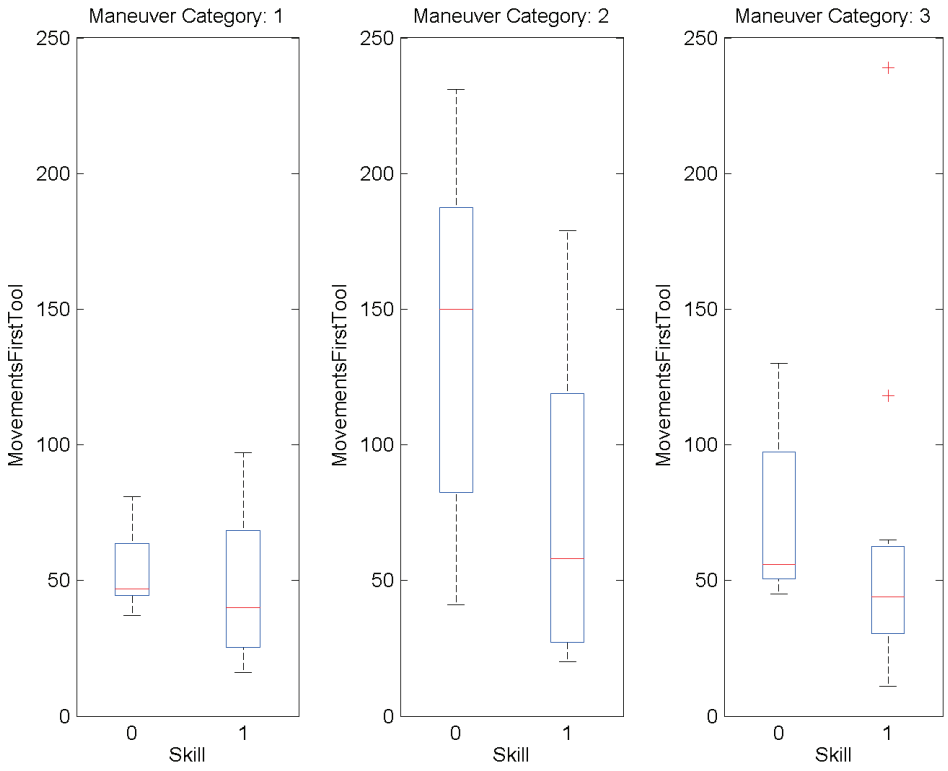


Figure 4.3: Number of movements in instrument motion to complete maneuvers in an interrupted suturing task. Task performances with low OSATS score (≤ 10) are marked as '0' on the x-axis, while those with a high OSATS score (≥ 27) are marked as '1'. Note: OSATS $\in [5, 30]$. © Springer International Publishing Switzerland 2014.

Challenges: No Groundtruth

Manual assessment of surgical skill for segments is impractical and of uncertain validity. While task-level manual assessment itself is challenging due to lack of rater (surgeon) time, assessing and assigning a score to each of the multiple segments within the task can easily become a resource-intensive effort. In addition, our attempts to manually assess surgical skill for maneuvers using GRS indicate that maneuvers may contain insufficient information to make an overall assessment of segment-level skill using current GRS tools. To our knowledge, no existing reliable and valid tools exist to manually assign GRS-like scores for segments. This makes development of automated tools and their validation for segment-level evaluation challenging as there is no groundtruth to begin with.

Solutions: Crowdsourcing and Pairwise Comparisons

As noted previously in Sections 1.2.4 and 3.6, crowdsourcing has shown reliability and validity in biomedical data analysis, specifically in surgical skill assessment.[23,24,173] These works showed that crowdsourced technical skill assessment scores using the GEARS tool were highly efficient and valid compared to those assigned by experienced surgeons. However, it has proven difficult to perform absolute assessment of segment-level skill.

On the other hand, pairwise comparisons[174] have been shown to yield valid assessments when absolute assessment is difficult – examples include assessing disease

severity,[175] movie recommendations,[176] ranking chess players,[177] ranking video game players[178] and information retrieval.[176,179] Intuitively, relative rating does not require a deeper understanding of the ordinal scale that is under survey compared to absolute rating. In the absolute rating scenario, if the rater hasn't seen the best and the worst samples out of the population, there is a lower chance of assigning the lowest/highest rating to the sample. However, in the relative rating scenario, comparing two samples and selecting one as better than the other doesn't require such exploration of the population extrema by the rater. We will refer to this selection of one given sample over the other as *preference*.

We believe that such preferences obtained from crowdsourced pairwise comparisons may provide efficient, reliable, and valid solutions for objective assessment of segment-level surgical technical skills.

In the remainder of this chapter, we will describe our novel approach for segment-level skill assessment using such pairwise comparisons-based preferences. We will present our pilot and validation studies on crowdsourcing and predicting segment-level scores for technical skills. We will show results of an extended validation of our framework in a cross data set testing setup with virtual reality and bench-top simulation training performances. Finally, we will address the question – absolute versus relative ratings, which is more reliable for segment-level skill assessment? – showing results from a crowdsourced comparison study.

4.2 Framework

Our skill assessment framework consists of three components as shown in Figure 4.4. The first component is an automated classifier to assign skill-based preferences in pairwise comparisons of task segments. We then use this classifier to compute percentile scores for task segments as an objective measure of segment-level skill. Finally, we compute an OSATS-like score for the overall task using the segment-level percentile skill scores.

4.2.1 Preference Classifier

The first component in our framework is a binary classifier that assigns a preference for a given pair of segments. We denote the preference relation using the symbols $\prec$ and $\succ$ and define it as follows:

$$\begin{aligned} m_1 \prec m_2 \qquad & \text{if } m_2 \text{ is better than } m_1 \\ m_1 \succ m_2 \qquad & \text{if } m_1 \text{ is better than } m_2 \end{aligned} \tag{4.1}$$

where, m_1 and m_2 are task segment performances (typically belonging to the same segment category). We believe performances close to one another in skill levels still have minor differences and preference relation will exist in such cases, although with a relatively low level of confidence.

Figure 4.4: Components of our framework (shown in pink blocks) for objective surgical skill assessment: 1) Preference Classifier, 2) Percentile Segment-level Scores, and 3) Overall Task Score. (a) The set R represents manual preferences assigned to pairs of segments in the set P by the raters. (b) Given a new performance of a task T, our framework assigns percentile scores to the constituent segments by comparing them against a library of performances L. An overall task-level score S_T is computed using the segment-level scores.

Based on this definition of preference, a binary classifier C is described as:

$$C(\mathbf{f}_1, \mathbf{f}_2) = \begin{cases} 1 & \text{if } m_1 \succ m_2 \ , \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

where, $\mathbf{f}_i$ is a feature vector representing the segment-level performance m_i using metrics for surgical skill. In general, the metrics or the feature representation can be computed from any surgical data modality. In our case, we use simple quantitative metrics listed in Table 4.2 (please refer to Section 2.3 for further details) derived from surgical motion data. The input feature vector for C is a concatenation of $\mathbf{f}_1$ and $\mathbf{f}_2$. We train the classifier using manually assigned pairwise preferences as the groundtruth labels.

4.2.2 Percentile Scores for Task Segments

The second component of our framework computes an objective skill score for individual task segments. Let us consider a surgical task that can be composed of a sequence of segments belonging to K different categories. For better legibility and ease of notation, let us look at segments within this task that belong to a single category (k) only. Now, let us define some notation:

- T is a task performance consisting of t segments,
- m_{T_j} is the j^{th} task segment belonging to the k segment category

Table 4.2: Quantitative metrics using motion data

Metric	Abbreviation	Description
Completion Time	CT	time in seconds to complete the task segment
Time Fraction	TF	fraction of overall task time spent in performing the segment
Path Length	PL	distance traveled by the instrument tip
Ribbon Area[112]	RA	area swept by the instrument shaft
Movements[53]	MV	number of peaks in magnitude of velocity of the instrument tip
Gripper Activations	GA	number of times the instrument gripper was closed
Working Distance	WD	distance between the instrument tip and endoscope tip along the view direction
Master Path Length[54]	MPL	distance traveled by the manipulators at the console of a tele-manipulation system
Master Workspace Volume[171]	MWV	bounding volume of the console range of motion

Refer to Section 2.3 for further details.

- $\mathcal{L} = \{m_{L_1}, m_{L_2}, \ldots, m_{L_n}\}$ is a corpus of n previously collected performances of k segment category

We apply C to compare m_{T_j} with all performances from the corpus $\mathcal{L}$ as shown in Figure 4.4b. Subsequently, we compute the percentile score (S_{T_j}) for m_{T_j} as follows:

$$S_{T_j} = \frac{1}{n} \sum_{i=1}^{n} C(\mathbf{f}_{T_j}, \mathbf{f}_{L_i}) \tag{4.3}$$

where, $\mathbf{f}_{T_j}$ and $\mathbf{f}_{L_i}$ are feature vectors corresponding to the segments m_{T_j} and m_{L_i}, respectively. The percentile score S_{T_j} for instance m_{T_j} is the proportion of pairwise comparisons between m_{T_j} and each instance m_{L_i} in $\mathcal{L}$ where $m_{T_j} \succ m_{L_i}$. A percentile score of 1.0 reflects that segment performance m_{T_j} is ranked above all the performances present in $\mathcal{L}$. Selecting $\mathcal{L}$ to be representative of the entire skill spectrum becomes important to make the percentile score a meaningful feedback.

4.2.3 Overall Task Score

The third component of our framework computes an objective measure of surgical skill for the overall task based on percentile scores of the constituent task segments. We hypothesize that a linear summation of the percentile scores for all segments within a task will yield an objective and valid overall task score. Accordingly, we train a linear regression model to learn parameters for each segment-level score in a task using expert-assigned global rating scores (GRS) as the ground-truth. Note, number

of segments t will vary across performances T. Thus, we quantize the percentile scores for task T as a vector $P_T \in \mathbb{R}^K$ (K is the number of segment categories). Each value in this vector is computed as the mean of percentile scores assigned to all segment occurrences of the category.

$$P_{T_k} = \frac{\sum_{j=1}^{t} S_{T_j}}{\sum_{j=1}^{t} 1} \qquad \text{such that} \quad T_j \in k \text{ segment category} \tag{4.4}$$

The linear model for overall task score is described below:

$$S_T = \beta_0 + \sum_{k=1}^{K} \beta_k \, P_{T_k} + \beta_e \, \mathbf{e}_T \tag{4.5}$$

where, S_T represents groundtruth GRS for T. We include $\mathbf{e}_T$ to account for the fraction of total task time spent in portions of the task which did not constitute a semantically meaningful activity segment. Previous work in literature by D'Angelo et al.[180] shows idle time in performance to be a valid measure of surgical experience and motor planning. We learn the coefficients β using a training set containing groundtruth GRS scores.

4.3 Pilot Study

Results from this study has been published previously: A. Malpani, S. S. Vedula, C. C. G. Chen, and G. D. Hager, "Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task" in Information Processing in Computer-Assisted Interventions, Springer International Publishing, 2014, pp. 138147.[157]

The goal of this study was to conduct a pilot validation of the hypothesis that a valid preference classifier can be trained using pairwise preference annotations and motion metrics. A secondary aim was to test the validity of the overall task score against expert surgeon assigned GRS scores. A tertiary aim was to collect pilot data on crowd's inter-rater reliability and validity to design a larger crowdsourcing study.

4.3.1 Data Set

We used the MultiSite Suturing data set described in Appendix A.1 for the surgical performance data. For our experiments described below, we used 502 maneuvers (out of a total of 1008) belonging to the four categories (ST1, ST2, KT1 and KT2). We omitted all the incomplete maneuver segments as well.

4.3.2 Framework Implementation

A smaller set of performance metrics (Table 4.2) were used to build the feature vector for each segment performance.

$$\mathbf{f} = [CT, PL_1, PL_2, RA_1, RA_2, MV_1, MV_2] \tag{4.6}$$

where, the subscripts indicate the instrument index. We used a linear support vector machine (SVM) as the binary preference classifier C.

We did not have a separate data corpus $\mathcal{L}$. For a given maneuver, we chose all

remaining maneuvers in the data set belonging to the same category as the corpus.

4.3.3 Preference Annotations

We recruited people from our project group to participate in this pilot study. An email describing the protocol as well as requesting their participation was sent out including instructions on how to complete the task. The experiment involved watching 80 pairs of maneuver videos (20 from each maneuver category, selected at random) and entering a binary response indicating which of the two videos in a each pair was performed with higher skill. The participants could flag their responses as being confident or not. We did not specify any explicit skill criteria for the annotation. Within two to three days, responses from 7 individuals had been obtained. We will refer to these set of 80 annotations as A_C. One of these participants annotated an additional 284 pairs of maneuvers. These maneuvers were randomly and uniformly sampled from across the four maneuver categories. We will refer to these set of 284 annotations as A_I. This study was conducted as a human subjects research (HIRB00001603) approved by the Johns Hopkins University Homewood IRB.

4.3.4 Validity and Reliability

We measured inter-rater agreement over the A_C preference annotations using Fleiss' kappa (Section 2.5). We measured the inter-rater agreement between the

expert surgeon and each non-surgeon member of the crowd using Cohen's kappa and percent agreement (Section 2.5).

We trained a separate SVM classifier for each maneuver category using the 364 annotations ($A_I + A_C$) of preferences – 264 annotated in A_I and part of A_C annotated by the same individual who annotated A_I. We tested the validity of the preference classifier using a cross-validation experiment by leaving out 30% of the annotations for testing. We repeated this 30 times.

Next, we again trained a separate SVM classifier for each maneuver category using all 284 pairs of maneuvers in A_I and tested on the 80 pairs of maneuvers annotated in A_C. We chose to do so to account for any bias in A_I (since these set of annotations were obtained from a single individual). We assessed validity of preferences assigned by the classifier by computing the accuracy compared against A_C for each member in the crowd. We used only the subset of preferences that the individuals marked as being confident.

We also assessed validity of overall task scores generated using our framework. We performed a leave-one-task-out validation. We used the combined set of $A_I + A_C$ annotations as before. We excluded all maneuvers m_{T_j} belonging to the held out task performance T from the training data. If m_{T_j} was one of the performances in a pair then that pair's preference annotation was excluded from the training set. Percentile score for m_{T_j} was computed using this specific classifier. This process was repeated for all the 135 task performances in the data set. Finally, a linear regression model

was obtained using segment percentile scores from the 134 task performances other than T and their respective GRS scores. This model was used to regress a GRS score for T. We computed a Spearman correlation coefficient between the predicted scores for the trial and expert assigned GRS.

4.3.5 Results

We observed moderate inter-rater agreement among manual preference annotations provided by the crowd (Fleiss' $\kappa = 0.42$, 95% CI = 0.40 to 0.43). The agreement was high when only preference annotations marked as being confident were compared (Fleiss' $\kappa = 0.88$, 95% CI = 0.85 to 0.91). We observed moderate to high agreement between the single expert surgeon and rest of the members of the crowd on average (84.36% for annotations marked as confident and 74.38% for all annotations in A_C). Table 4.3 shows the inter-rater agreement between the expert and each member of the crowd. The crowd also indicated that there were pairs of performances in the experiment, which were similar in skill, making it hard to decide which one of them was better. However, there was only a low agreement (Fleiss' $\kappa = 0.17$) on deciding whether the presented pair was confusing amongst the crowd. Similarly, we compared the percent agreement between the annotator of A_I and the crowd. The percent agreement between the remaining six crowd members was 83.27%, while the average percent agreement of the A_I annotator against each of the crowd members was 84.43%.

Table 4.3: Inter-rater agreement between expert and each crowd member

		Crowd Member					
		1	2	3	4	5	6
# of pairs		56	48	43	51	44	32
Cohen's κ	all	0.63	0.53	0.39	0.43	0.45	0.49
	conf	0.72	0.66	0.72	0.73	0.53	0.75
Perc Agr	all	81.25	76.25	70.00	71.25	72.50	75.00
	conf	85.71	83.33	86.05	86.27	77.27	87.50

all: refers to 80 pairs in A_C,
conf: refers to annotations marked as confident by expert and the crowd member

Table 4.4: Accuracy of maneuver-specific preference classifiers using a 30-fold held-out validation setup

	ST1	ST2	KT1	KT2
Mean (std)	85.71 (6.91)	82.92 (6.95)	90.23 (3.52)	80.06 (5.88)

Combined set of annotations $A_C + A_I$ was used for this experiment.

Our initial cross-validation revealed that the trained classifier using data from $A_I + A_C$ could automatically annotate preferences with an average accuracy of at least 80% (Table 4.4). Although agreement among manual annotators was lower when we used all preference annotations (irrespective of being marked confident), the classifier's annotations were still consistent with those from the crowd. We observed moderate accuracy while comparing the preference predictions by the preference classifier against annotations provided by each member of the crowd (Table 4.5).

Finally, we found overall trial-level scores obtained from the linear regression model were moderately correlated with manually assigned GRS (Figure 4.5). The Spearman's correlation coefficient was 0.47 ($p < 0.001$).

Table 4.5: Accuracy of preference classifier against members of the crowd

	Member						
	1	2	3	4	5	6	7
confident	75.00	85.71	76.36	82.35	68.25	69.23	80.00
all	68.75	77.5	65.00	76.25	62.50	71.25	71.25

Training data was A_I, Test data was A_C from each crowd member.
confident: refers to A_C annotations that were marked as confident.
all: refers to A_C.

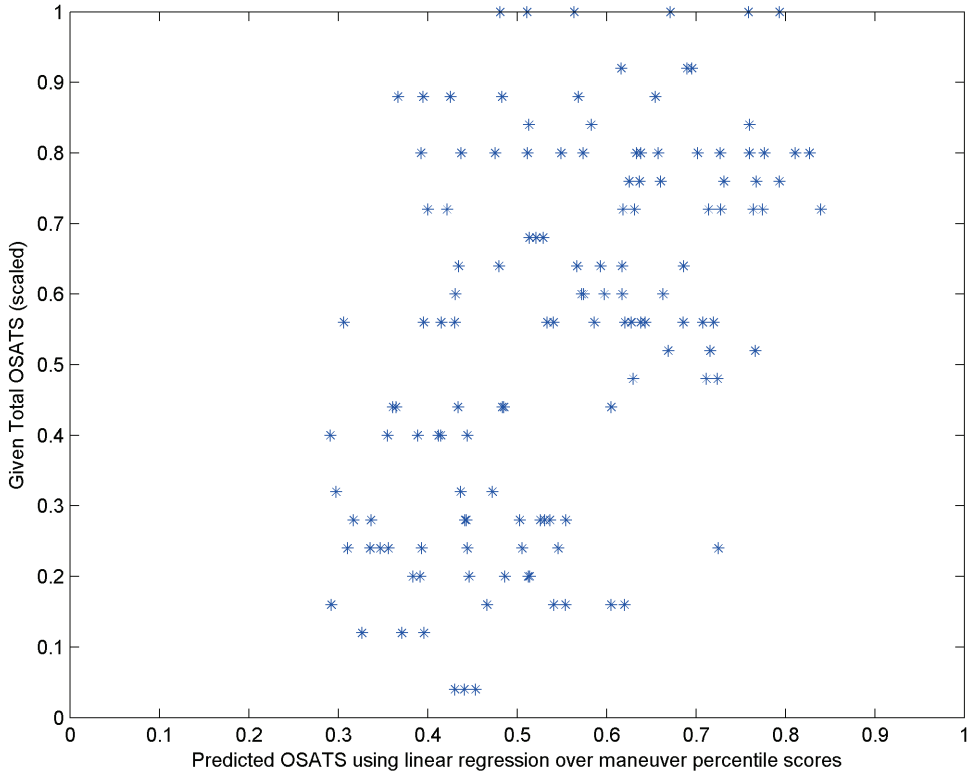


Figure 4.5: Plot of predicted task scores using a linear regression over maneuver percentile scores generated by our framework. The manual GRS using OSATS have been scaled to the [0,1] interval on the y-axis.

4.3.6 Outcomes and Future Work

Our analyses based on an interrupted suturing task data set show that a reliable and valid preference classifier can be trained to automatically annotate the better-performed maneuver in a given pair. Thus, we can efficiently generate such pairwise comparisons in an automated fashion for large data sets.

Sensitivity of the annotations to the constitution and characteristics of the crowd remains an open question. The constitution of the crowd may affect performance of our framework and should be investigated in future research.

4.4 Validation Study

In the previous pilot study, using a limited sample, we demonstrated that crowdsourcing can yield reliable and valid pairwise comparison of surgical skill at the segment-level. In this study, we extend our analysis with a larger sample size, and also explore the computation and validation of overall task scores using segment-level percentile scores. The goals of our study were: 1) to establish reliability and validity of a framework to objectively assess surgical skill using pairwise comparisons of task segments, and 2) to compare assessments obtained from our framework using pairwise comparisons from two sources - a surgically untrained crowd and a group of expert surgeons.

Results from this study have been previously published in a journal article: A. Malpani, S. S. Vedula, C. C. G. Chen, and G. D. Hager, "A study of crowdsourced segment-level surgical skill assessment using pairwise rankings" Int J CARS, vol. 10, no. 9, pp. 14351447, Jun. 2015.[115]

4.4.1 Data Set

We used the MultiSite Suturing data set described in Appendix A.1. Compared to the previous study we used all the maneuver categories *viz.* ST1, ST2, GPR, KT1 and KT2. The previous work omitted IMS and incomplete maneuvers. In this study, we accounted for these using the $\mathbf{e}_T$ factor in the linear regression model (Equation 4.5).

4.4.2 Preference Annotations

We conducted a crowdsourcing user study (HIRB00001603 approved by Johns Hopkins Homewood IRB) to generate two different sources of groundtruth[1] for training the preference classifier in our framework – surgically untrained individuals (crowd) and faculty surgeons (experts). We hosted a survey on a website for the crowd and expert participants to complete the specified HITs [2], which in our case was to provide preferences for pairs of maneuver performances. The study call was voluntary and open to all within the Johns Hopkins community. Each annotation task (AT) was created using pairs of maneuvers belonging to the same category. We did not include IMS when generating ATs because the actions performed across instances of IMS in our data set were highly variable in nature, and in the goals they accomplished. However, we did account for the time spent in IMS while predicting overall task scores.

[1]The term groundtruth, here and henceforth, has been used here to denote a *reference* value obtained by pooling the crowd/expert responses.

[2]Refer to Section 2.4 for details on crowdsourcing terminology

The maneuver videos were typically 20 to 30 seconds in length.

Based on a priori sample size calculations, we sampled a total of 360 ATs for the crowd and a subset of 120 of those 360 ATs for the experts. We assumed that the proportion of pairs with correct ordering will be 85% for the crowd and 90% for the experts. Accordingly, we computed that we will be able to estimate the proportion of pairs with accurate ordering of videos with a 95% confidence interval (CI) of width of 0.1 (10%) if we recruited 49 crowd participants and 35 expert participants. Furthermore, we computed the sample size to test a hypothesis of equivalence comparing accuracy of the preference classifiers trained using preferences obtained from crowd and expert participants. We assumed that the accuracy of classifier trained with crowd ratings will be 80% and accuracy of classifier trained with expert ratings will be 85%. We estimated that we would have 90% power to establish equivalence within a 10% margin with 52 unique pairs of videos.

We grouped ATs into 12 HITs of 30 ATs each for crowd participants, and two HITs of 30 ATs and six HITs of 15 ATs for experts. This division satisfied the required sample size while making the overall length of the surveys shorter to encourage expert participation. A study participant was required to complete all the ATs belonging to a HIT for their participation to be complete. Additionally, attention ATs consisting of an obviously good performance versus an obviously poor performance were presented to the participants at regular intervals (every 10 ATs). Participants who didn't provide correct preferences for such ATs were automatically disqualified

from the study.

The participants were displayed an informed consent for the study on the welcome page and were registered using their name and Hopkins email address. The participants could participate in any number of HITs, but only once per HIT. Participants who failed an attention AT were not allowed to participate in any other subsequent HIT. The crowd HITs had a reward of $10 gift card and a duration limit of three days. The expert HITs had no reward and a duration limit of seven days. There were no restrictions on the amount of time spent by a participant on an individual AT.

Figure 4.6 illustrates a typical screen visualized by study participants during an AT. We asked the participants to specify which of the two maneuvers displayed on the screen appeared to have been performed with greater skill (preference), and to specify their level of confidence in choosing the preference (as shown in Figure 4.6) on a Likert-like scale. The answer options were enabled, only when the participant had viewed both the videos completely.

We recruited 147 crowd participants across the 12 HITs, most were students from the engineering, arts and sciences programs at Johns Hopkins University, a few were staff members. We restricted the total number of crowd assignments per HIT (HIT 1: 52 assignments, HITs 2 through 9: 11 assignments, HITs 10 through 12: 5 assignments). We recruited eight expert participants across the eight HITs, all of whom were faculty surgeons at the Johns Hopkins Medical Institutions. We restricted the recruitment to three experts per HIT to get multiple responses for each of the 120 ATs

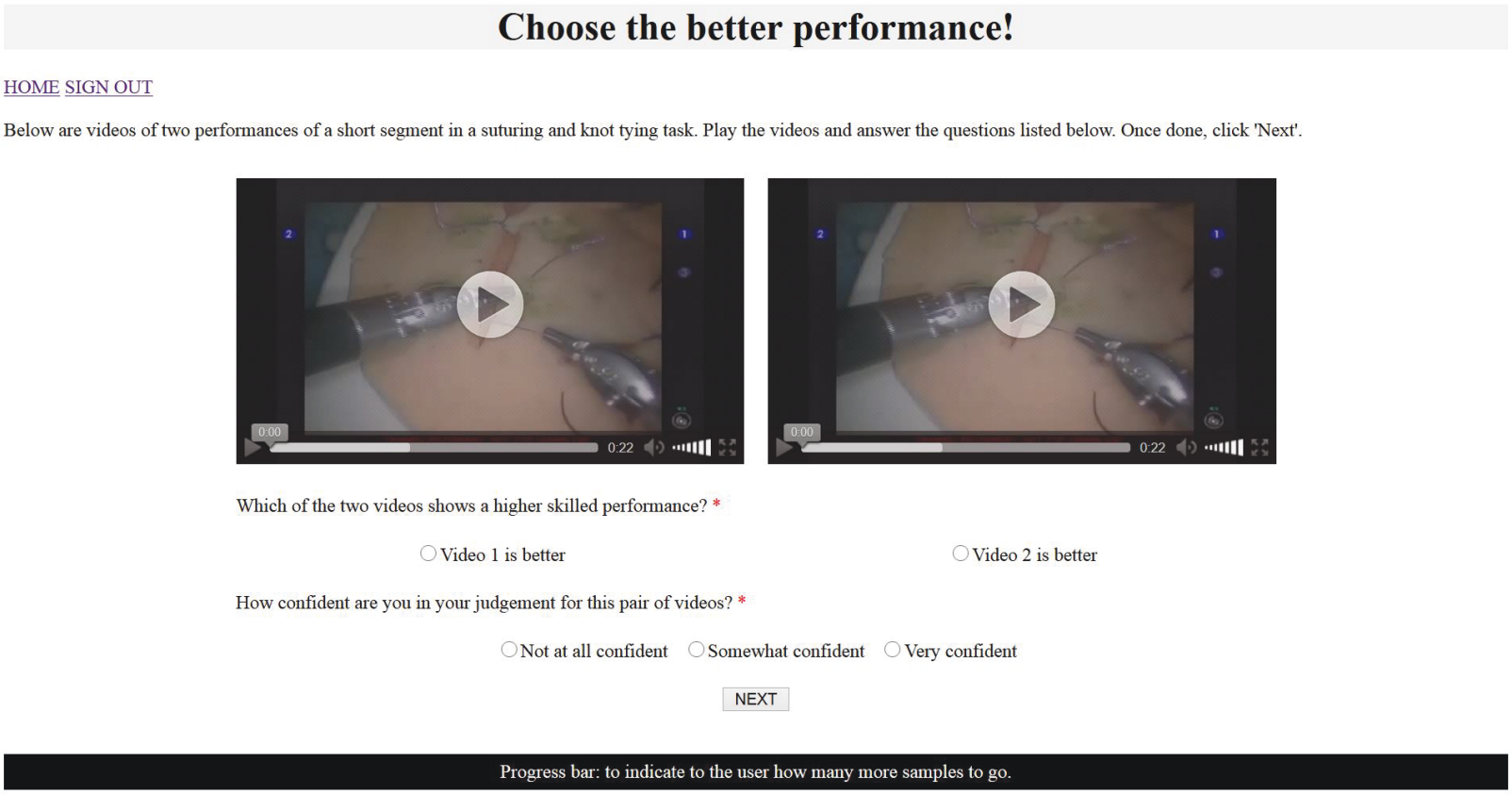


Figure 4.6: A snapshot of the web-based HIT page showing a sample preference annotation task (AT). © CARS 2015

sampled. We obtained preferences from all the crowd participants within a period of three days, whereas it took about four weeks to capture preferences from the experts. For this reason, we were not able to recruit the number of experts suggested by our power analysis, although, as we note later, the consistency of the experts suggests our analysis was overly conservative. The time spent (in seconds) per AT [3] across the 120 overlapping ATs were: experts (mean: 117.36, σ: 230.52), and crowd (mean: 71.52, σ: 87.91).

4.4.3 AT Agreement and AT Confidence

For each AT, we computed two properties *viz.* percent agreement and confidence. Percent agreement (agr) was computed as in Section 2.5, however, using responses which were marked with a confidence level of five only. To ensure that our preference classifier was trained on a meaningful ground truth, we used only those ATs for training where $agr \geq 0.75$.

The second characteristic property of the ATs is the confidence ($conf$), which was computed as an average of confidence level weights (Table 4.6) assigned by participants responding to that AT as shown in Eq. 4.7.

$$conf_t = \frac{1}{k_t} \sum_{j=1}^{k_t} w_{tj} \tag{4.7}$$

[3] The participants could take breaks and come back and answer these ATs later. Thus, we cannot draw any reliable conclusions based on these numbers.

Table 4.6: Confidence levels elicited in the HIT and corresponding weights for ratings

Survey Phrase	Level	Weight
Very confident	5	1.0
Somewhat confident	3	0.5
Not at all confident	1	0.0

where, w_{tj} is the confidence weight (Table 4.6) associated with the confidence level indicated by the participant j for their preference for the AT t, and k_t is the total number of assignments for the AT t. By doing so, the classifier was trained using data where the raters were more confident about their preferences. We used HITs with $conf \geq 0.5$ in our sensitivity analysis.

4.4.4 Pooling Preferences

To obtain a single ground truth preference per AT (pair of segments), we investigated three different approaches for majority pooling and one approach for weighted pooling.

In the first approach, we simply selected the majority rating (R_{all}) from all the preference ratings obtained for a given AT. In the second approach, we selected the majority among ratings where the confidence level was at least three (R_3). In the third approach, we selected the majority among ratings where the confidence level was five (R_5). We used all three approaches for reliability analyses, but only the simple majority rating approach (R_{all}) for validity analyses due to sample size limitations

with the remaining majority pooling approaches.

In the weighted pooling approach, we selected the preference using a weighted count of preference ratings for a given AT (R_w). Table 4.6 shows the weights we used for each level of confidence associated with the ratings. Ratings with confidence level 5 contributed a full count and those with confidence level 3 contributed one-half of a count towards the preference ratings. Ratings with confidence level 1 did not contribute to the preference rating in this weighted pooling approach.

4.4.5 Reliability and Validity Experiments

Now, we will present our approaches towards verifying the reliability and validity of (1) preference annotations, (2) preference classifier predictions, and (3) overall task score predictions.

Preference Annotations

We evaluated the inter-rater reliability of preferences separately for the crowd and experts using the Fleiss' kappa (κ) and percent agreement (*perc agr*) (refer Section 2.5). We also evaluated validity of preferences obtained from the crowd assuming preferences obtained from the experts were the groundtruth. We computed the percent agreement (*perc agr*) and Cohen's kappa (κ) as measures of validity (refer Section 2.5). Additionally, we computed the Fleiss' kappa for the confidence level ratings assigned by the crowd and expert participants. We compared the agreement

within crowd and expert groups in selecting the majority confidence rating using percent agreement.

Preference Classifiers

We used the entire set of metrics listed in Table 4.2 to build the feature vector shown in Eq. 4.8 for each maneuver performance.

$$\begin{aligned}\mathbf{f} = \big(&CT, TF, PL_1, PL_2, RA_1, RA_2, MV_1, MV_2, \ldots \\ &GA_1, GA_2, WD_1, WD_2, MPL_1, MPL_2, MWV_1, MWV_2\big) \\ &\in \mathbb{R}^{16}\end{aligned} \tag{4.8}$$

We trained two separate linear support vector machines (SVM), one using preferences from the crowd and the other from experts. We explored an AdaBoost classifier using stump-based weak learners as well. However, the SVMs performed better than the boosted classifier and thus further analyses were performed using SVMs.

We trained each of these SVMs using two different sets of features; the first set (SVM7) matched the 7-D feature vector used in the pilot study (Eq. 4.6), and the second set (SVM16) included the 16 dimensions described in Eq. 4.8. We trained a separate classifier for each maneuver category as well as one overall classifier for all categories of maneuvers pooled together. In addition, we trained separate classifiers for preferences obtained with two pooling approaches - R_{all} and R_w (Section 4.4.4).

We evaluated crowd- and expert-based preference classifiers against the respective

pooled preferences as the groundtruth. We used a 10-fold cross-validation approach and computed accuracy between the classifier-assigned preferences and participant-assigned pooled preferences. We computed the accuracy of the crowd preference classifier varying the number of training samples used. A fraction (20%) of the HITs was held out as a fixed test data set. The number of training samples (n) was incremented in steps of 10 samples at a time. For each n, an average accuracy was calculated using 20 bootstrap iterations for sampling the training data.

Overall Task Skill Scores

For the corpus $\mathcal{L}$ defined in Section 4.2.2, we used the same approach as in the pilot. For a given maneuver, we chose all remaining maneuver instances of the same category in the data set to form the corpus.

We compared the task-level scores obtained using the expert preference classifier against groundtruth GRS. We trained a simple linear regression model (Eq. 4.5) to predict the groundtruth GRS in a leave-one-out cross validation approach. The predictors for the model included the segment-level percentile scores as a four dimensional vector (ST, GPR, KT1, KT2), the number of IMS, fraction of total task time spent performing IMS, and the fraction of total task time that was not annotated with any maneuver label. The latter three terms in the predictors formed $\mathbf{e}_T$ from Eq. 4.5. The segment score for ST was obtained from the score for ST1 or ST2, whichever was performed in the given instance of the task.

We computed the root mean squared error (RMSE) and the Spearman's correlation coefficient (ρ) between predicted and groundtruth scores as measures of validity. The Spearman correlation is a non-parametric measure of association between two ranked variables. A value of $\rho = +1$ indicates perfect monotonic dependence, while a value of zero indicates no correlation. In addition, we learned similar regressions to predict scores for each of the six individual components within OSATS listed in Figure 2.14.

4.4.6 Comparison of Crowd and Expert Preference Classifiers

We assessed the crowd and expert preference classifiers for three outputs of our framework pipeline:

I. Accuracy: We tested the equivalence of the crowd and expert preference classifiers by checking whether the accuracy of the crowd preference classifier is within the 10% margin of accuracy of the expert preference classifier. For hypothesis testing purposes, we performed cross validation using the set of ATs rated by both the crowd and the experts ($n = 75$)[4], while training the respective classifiers using all held out data available per group of users (crowd had larger number of ATs). Additionally, we performed a sensitivity analysis using only those ATs rated by both the crowds and

[4]The number of ATs rated by both the crowd and experts was 120. However, filtering the ATs based on the agreement metric (Section 4.4.3) drops the count to 75.

experts for training as well as testing in a leave-one-out cross validation approach. More training data was available for the crowd classifier as compared to the expert classifier in the former analysis, whereas the training data for the two classifiers was the same in the latter case.

II. Segment-level Scores: We computed a Spearman's correlation coefficient (ρ) between the segment-level scores obtained from the crowd and expert preference classifiers, separately for each maneuver category.

III. Task-level Scores: We computed a Pearson's correlation coefficient (ρ) between the task-level scores obtained using the crowd and expert preference classifiers. The Pearson correlation measures the linear correlation between two continuous variables. A value of +1 for the Pearson correlation indicates total positive correlation, 0 indicates no correlation, and -1 indicates total negative correlation. In addition, we tested whether the task-level scores obtained using the crowd and expert preference classifiers were statistically equivalent to each other within a pre-specified margin of two units on the GRS scale.

4.4.7 Results

Reliability and Validity of Preference Annotations

As shown in Table 4.7, we observed moderate inter-rater agreement within both the crowd and expert participants. Experts appeared to have a higher inter-rater

Table 4.7: Inter-rater agreement for crowdsourced preferences

	Crowd	Expert
# of ATs	360	120
# of workers	147	8
perc agr (95% CI)	0.81 (0.80, 0.83)	0.88 (0.85, 0.91)
κ (95% CI)	0.41^{*} (0.40, 0.42)	0.55^{*} (0.45, 0.64)

perc agr: percent agreement, κ: Fleiss' kappa
* $p < 0.001$

agreement compared with the crowd as one would expect.

The crowd preferences were at least 83% accurate when taking expert preferences as the groundtruth. This accuracy was robust across all four approaches for pooling preferences (Section 4.4.4) for a given AT, as shown in Table 4.8. The accuracy increased with the R_3 and R_5 pooling approaches as one would expect with ratings having higher confidence.

Inter-rater agreement on choosing the confidence level seemed to be higher for ATs with higher majority confidence levels for both the crowd and experts, as shown in the Table 4.9. However, the overall inter-rater agreement (using Fleiss' kappa) on confidence level ratings was observed to be very low - 0.08 (crowd) and 0.22 (experts).

Validity of Preference Classifiers

A preference classifier trained using ratings obtained from the crowd could predict the crowd's pooled preferences with an accuracy of 85% (std. error: 2%). The preference classifier trained by expert preferences had an accuracy of 89% (std. error:

Table 4.8: Agreement between pooled preferences for ATs which were rated by both the crowd and expert participants

	Pooling Approach			
Statistic	R_{all}	R_3	R_5	R_w
# HITs	120	118[1]	89[2]	120
perc agr	0.83	0.85	0.87	0.84
95% CI	$(0.77, 0.90)$	$(0.78, 0.91)$	$(0.79, 0.94)$	$(0.78, 0.91)$
κ	0.66	0.69	0.73	0.68
95% CI	(0.49, 0.84)	(0.51, 0.87)	(0.52, 0.94)	(0.50, 0.86)

ATs that did not get ratings with confidence levels of at least 3[1] and at least 5[2] respectively, were omitted while computing the above statistical measures.
perc agr - percent agreement, κ: Cohen's kappa

Table 4.9: Inter-rater agreement for crowdsourced confidence levels

	Percent Agreement (%)	
Confidence Level	Crowd	Expert
1	49.89	53.47
3	57.70	72.78
5	65.22	84.80

3%). As noted before in Table 4.7, crowd participants' inter-rater agreement across the ATs was 81% (95% CI: $80, 83$), while the experts had an agreement of 88% (95% CI: $85, 91$). Thus, the performance of our classifier is above par compared to the inter-rater agreement.

The accuracy of the crowd preference classifier improved when the training data were filtered to only include ATs with an overall confidence of 0.5 or more (see Table 4.10). But this was not the case with the expert preference classifier, where the accuracy appeared to decrease when we filtered the training data to include ATs with an overall confidence of 0.5 or more. Extending the set of training features did not appear to consistently improve accuracy of either the crowd or expert preference classifier.

Accuracy of the preference classifiers did not appear to be sensitive to whether we pooled preferences using R_{all} or R_w. Accuracy for the expert preference classifier for R_{all} was consistently greater than those for R_w, but the difference was small in magnitude. We did not observe a consistent direction for these differences with the crowd preference classifier (see Table 4.10).

Table 4.10 also shows that classifiers specific to some maneuver categories (KT1 and KT2) appeared to be more accurate than the overall classifier in predicting manual preferences. This was not the case for classifiers specific to other maneuver categories (ST1, ST2, and GPR).

The average accuracy of the crowd preference classifier trained using a varying

Table 4.10: Accuracy values for preference classifiers with crowd and expert preferences.

(a) Preference Classifier trained using **Crowd** Preferences

Pooling	Segment	HITs ($agr \geq 0.75$)			HITs ($agr \geq 0.75, conf \geq 0.5$)		
		N	SVM7	SVM16	N	SVM7	SVM16
R_{all}	ST1	30	0.73 (0.08)	0.40 (0.09)	20	0.75 (0.10)	0.65 (0.11)
	ST2	53	0.74 (0.06)	0.70 (0.06)	46	0.76 (0.06)	0.78 (0.06)
	GPR	54	0.78 (0.06)	0.74 (0.06)	41	0.78 (0.06)	0.78 (0.06)
	KT1	62	0.92 (0.03)	0.85 (0.04)	60	0.92 (0.04)	0.85 (0.05)
	KT2	78	0.88 (0.04)	0.88 (0.04)	73	0.90 (0.03)	0.89 (0.04)
R_{all}	ALL	277	0.81 (0.02)	0.82 (0.02)	240	0.82 (0.02)	0.85 (0.02)
R_w	ALL	277	0.81 (0.02)	0.80 (0.02)	240	0.85 (0.02)	0.86 (0.02)

(b) Preference Classifier trained using **Expert** Preferences

Pooling	Segment	HITs ($agr \geq 0.75$)			HITs ($agr \geq 0.75, conf \geq 0.5$)		
		N	SVM7	SVM16	N	SVM7	SVM16
R_{all}	ST1[1]	–	–	–	–	–	–
	ST2	15	0.47 (0.13)	0.80 (0.10)	14	0.50 (0.13)	0.71 (0.12)
	GPR	20	0.85 (0.08)	0.75 (0.10)	20	0.90 (0.07)	0.75 (0.10)
	KT1	25	0.88 (0.06)	0.92 (0.05)	23	0.87 (0.07)	0.87 (0.07)
	KT2	26	0.92 (0.05)	1.00 (0.00)	24	0.83 (0.08)	0.96 (0.04)
R_{all}	ALL	89	0.89 (0.03)	0.89 (0.03)	84	0.85 (0.04)	0.86 (0.04)
R_w	ALL	89	0.87 (0.04)	0.87 (0.04)	84	0.83 (0.04)	0.85 (0.04)

Values in parentheses are standard errors
Training data was filtered using *agr* and *conf* values for ATs defined in Section 4.4.3
N is the number of HITs available for cross validation after the filtering
SVM7 was trained using a subset of metrics (Eq. 4.6)
SVM16 was trained using all the metrics (Eq. 4.8).
[1] N was too low to perform cross validation

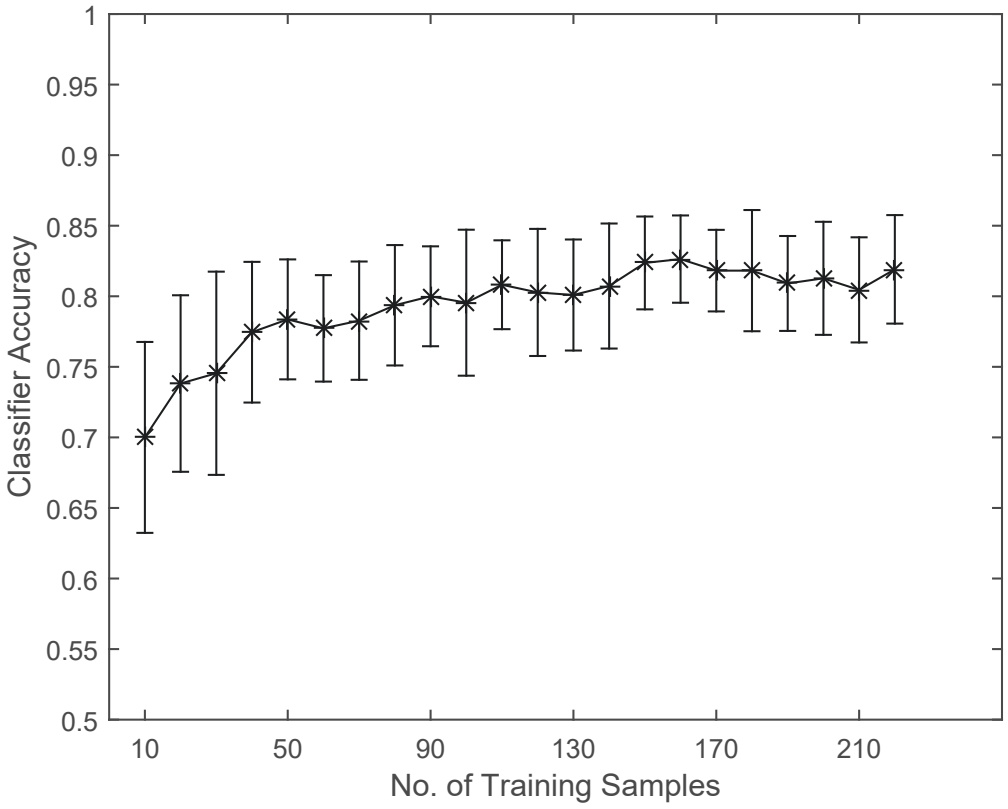


Figure 4.7: Crowd preference classifier accuracy v/s the number of training samples. The points on the plot are mean accuracy over a bootstrap sampling of 20 iterations for each setting of the number of training samples. The error bars indicate the standard deviation in the accuracy over these 20 iterations. © CARS 2015.

number of training samples is shown in Figure 4.7 and listed in Table 4.11. The accuracy plateaus after $n = 120$ training samples with a value of 0.80 showing a change in the order of 0.02 as the number of training samples varies in the range of $(120, 220)$. We did not conduct a similar analysis for the expert preference classifier due to a small sample size.

Table 4.11: Preference classifier accuracy for different number of training samples

# of samples (N)	Accuracy (%)	Standard Deviation (%)
10	70.00	6.76
50	78.36	4.25
90	80.00	3.54
130	80.09	3.93
170	81.82	2.89
210	80.45	3.72

Table 4.12: Validity of predicted task scores v/s expert-assigned OSATS

OSATS component	RMSE	ρ
Overall	5.54	0.55
Respect for tissue	1.05	0.52
Time and motion	0.95	0.56
Instrument handling	1.16	0.53
Knowledge of instruments	1.01	0.63
Flow of operation	1.20	0.45
Knowledge of specific procedure	1.14	0.33

RMSE: root mean squared error, ρ: Spearman's correlation
All correlations had $p < 0.001$

Validity of Overall Task Skill Scores

Using the expert preference classifier, we predicted expert-assigned overall GRS with RMSE lower than one standard deviation (σ) of the ground-truth (RMSE = 5.54; 0.85 σ). The Spearman's correlation between the predicted and ground-truth GRS was 0.55 ($p < 0.001$).

We predicted the individual components within OSATS as well. Results are shown in Table 4.12.

Comparison of Crowd and Expert Preference Classifiers

I. Accuracy: As shown in Figure 4.8, our analyses did not demonstrate equivalence between the crowd and expert preference classifiers within a margin of 10%. Our observation is consistent for SVM7 and SVM16 using training data obtained with different pooling approaches and filtered based on confidence property of the ATs. Using the same training data for the crowd and expert preference classifiers did not alter the outcome of the analysis, as can be seen in Figure 4.9.

II. Segment-level Scores: In the case of SVM7, segment-level scores obtained using the crowd preference classifier were highly correlated with those from the expert preference classifier ($\rho \geq 0.86$ for all maneuver categories; Spearman's). But in the case of SVM16, the correlation between the segment-level scores from the two preference classifiers was very sensitive to the sample size specific to the maneuver category. The correlation coefficient was as low as 0.11 for ST1 ($N = 30$) and as high as 0.85 for KT1 ($N = 62$).

III. Task-level Scores: Task-level scores predicted using segment-level scores from the crowd preference classifier were also highly correlated with those from the expert preference classifier ($\rho \geq 0.84$; Pearson's). As shown in Figure 4.10, task-level scores obtained using the crowd preference classifier were statistically equivalent to those obtained using the expert preference classifier within a margin of two units on the GRS scale.

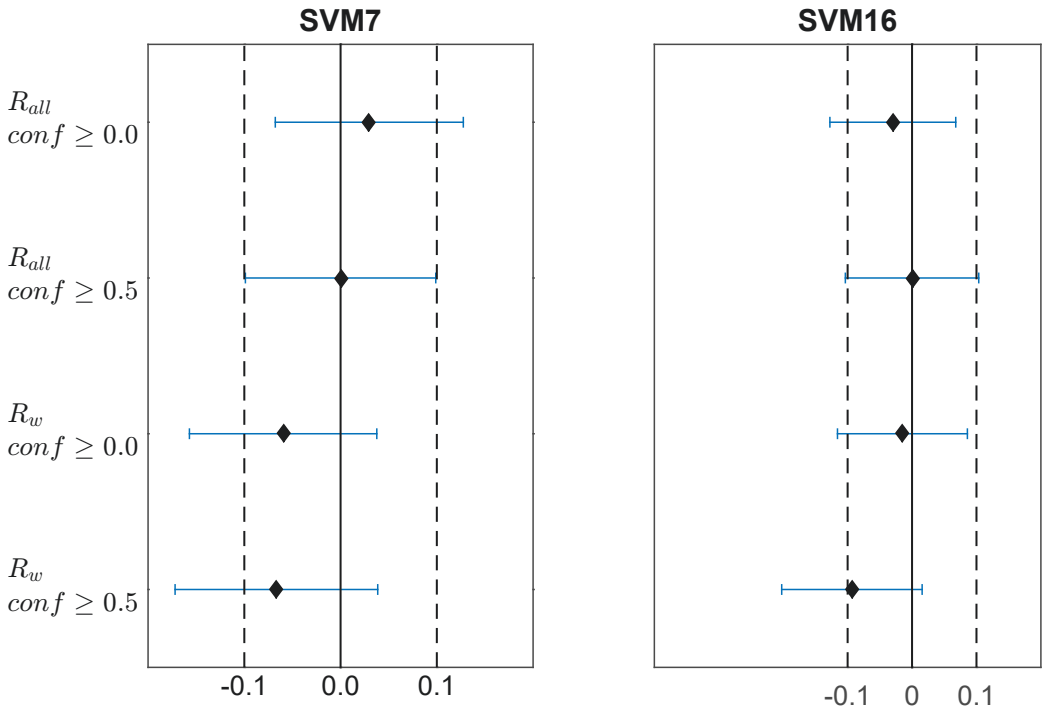


Figure 4.8: Equivalence testing of accuracy of crowd and expert preference classifiers using all available training data. The X-axis is the difference in accuracy from the crowd and expert preference classifiers. The dashed lines illustrate the equivalence margin on either side of the null value (solid line). The solid diamonds represent the estimate of the difference in task-level scores obtained from the two classifiers. The horizontal bars are the 95% CI for the estimates. Equivalence holds if the 95% CI lie entirely within the region bounded by the dashed lines. © CARS 2015.

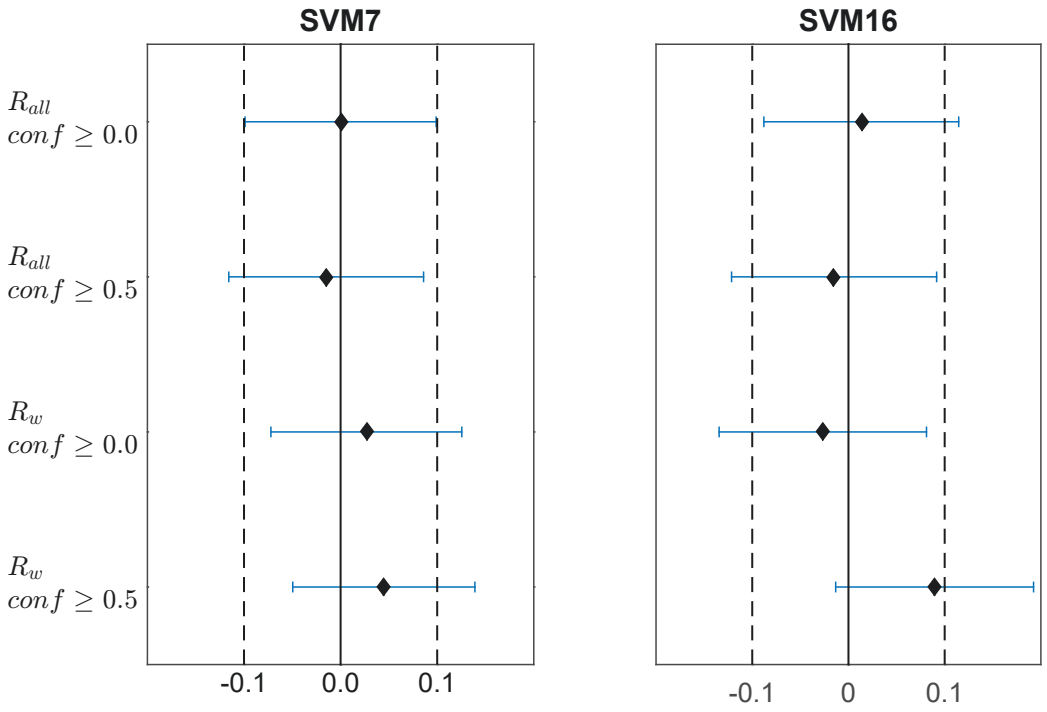


Figure 4.9: Equivalence testing of accuracy of crowd and expert preference classifiers using common training data. The X-axis is the difference in accuracy from the crowd and expert preference classifiers. The dashed lines illustrate the equivalence margin on either side of the null value (solid line). The solid diamonds represent the estimate of the difference in task-level scores obtained from the two classifiers. The horizontal bars are the 95% confidence intervals (CI) for the estimates. Equivalence holds if the 95% CI lie entirely within the region bounded by the dashed lines.

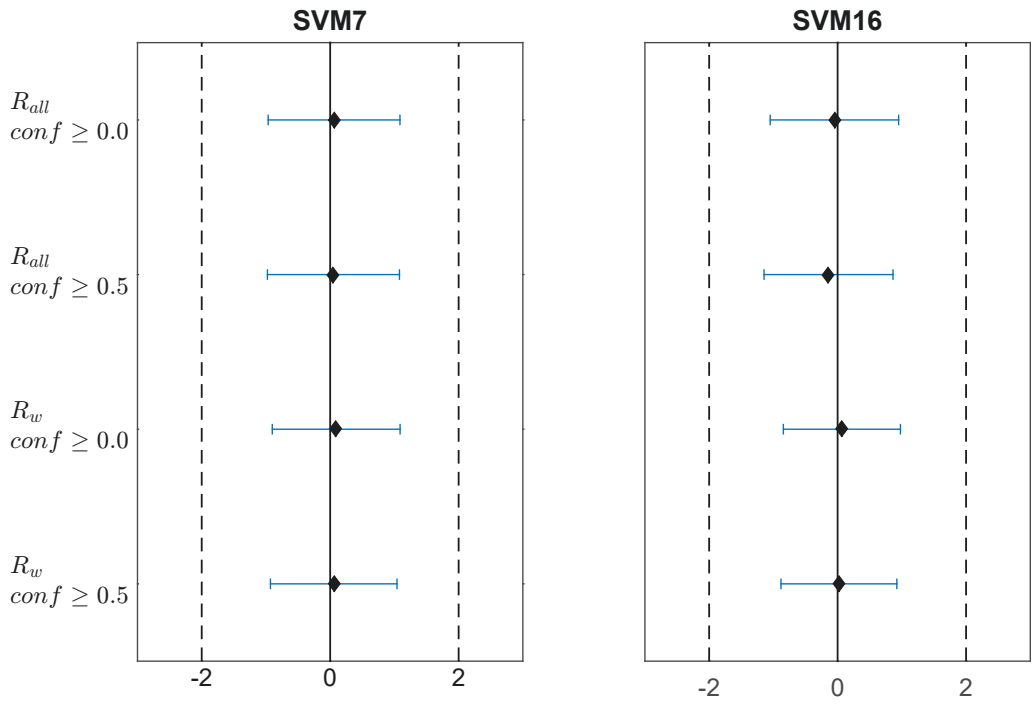


Figure 4.10: Equivalence testing of overall task scores from crowd and expert preference classifiers using common training data. The X-axis is the difference in overall task scores from the crowd and expert preference classifiers. The dashed lines illustrate the equivalence margin on either side of the null value (solid line). The solid diamonds represent the estimate of the difference in task-level scores obtained from the two classifiers. The horizontal bars are the 95% confidence intervals (CI) for the estimates. Equivalence holds if the 95% CI lie entirely within the region bounded by the dashed lines. © CARS 2015

4.4.8 Outcomes

Our findings in this study are strongly supportive of our framework for objective surgical skill assessment using pairwise comparisons of task segments. Our data indicate that assessments of segment-level skill can be obtained with moderate reliability from both surgically untrained individuals and from expert surgeons. Further, we show that crowdsourcing is an efficient, reliable, and valid solution for assessing surgical skills at the segment-level. The crowd yielded preferences for maneuvers with high validity when compared with expert surgeons (Table 4.8), and within three days compared with about four weeks for experts. The experts in our sample were affiliated with various surgical divisions and represented a wide range of experience (number of years in practice). Given the agreement among these diverse experts that we observed in our sample, we expect that our findings will be robust to ground truth specified by a larger group of experts.

Accuracy of manual preferences by the crowd translated directly into validity of all aspects of our framework. Given ground truth pairwise preferences for task segments, we demonstrated that a classifier can be trained with sufficient accuracy to yield valid and objective skill assessments at both the segment- and task-levels (Table 4.10). We did not observe a consistent improvement in accuracy of the preference classifier by extending the set of features from SVM7 to SVM16. Even though the accuracy for the crowd and expert preference classifiers were not equivalent, both segment- and task-level scores obtained from the two classifiers were highly comparable (Figures

4.8, 4.9 and 4.10). Furthermore, our framework yielded task-level GRS with an error that is comparable in magnitude to the variability we observed in our data set for task-level GRS assigned by an expert surgeon.

Our study establishes a basis for evaluating the educational value of targeted feedback based upon segment-level skill assessment.

4.4.9 Future Work

We have studied a single surgical task, interrupted suturing performed on using dVSS. Further studies validating our framework may focus on other tasks within typical surgical skills training curricula performed and perhaps using non-robotic surgical platforms.

Our results show that a single preference classifier trained over all available maneuver categories performed equally well compared to maneuver-specific training. An important next step after such a validation is testing generalization of the preference classifier from one surgical task to another with some shared notion of skill.

True validity of the framework relies on the availability of a corpus against which such pairwise comparisons will generate meaningful and useful skill scores.

4.5 Cross Data Set Validation Study

In the previous studies, we tested the reliability of the crowd in generating valid preferences on pairwise comparisons of maneuvers performed during bench-top suturing tasks. We showed the validity of a binary preference classifier trained using the crowd annotations in predicting preferences, generating segment skill scores and overall task scores. We hypothesize that such a preference classifier can be trained in the VR training setting as well using the same pairwise comparisons strategy. We also postulate that skill assessment based on pairwise comparisons will be valid across platforms, e.g. bench-top versus VR. Our goal in this study was to validate a framework for segment-level skill assessment in VR and to establish its cross-platform validity.

4.5.1 Data Set

We used a data set of 30 performances of the Suture Sponge I task (Figure 4.11) performed on a da Vinci Skills Simulator™ captured in a previous study.[181] The instrument motion and events data from the da Vinci API (Section 2.1.3) were captured along with the stereo video of the simulation. Suture Sponge I is a basic needle passing task. A slightly deformable sponge (yellow rectangular object in Figure 4.11) is in the center of the virtual workspace. A needle is placed on the top surface of the sponge at the start. Two virtual large needle driver instruments are provided to perform the

Table 4.13: Needle passing configurations in the suture sponge task

Maneuver	Driving Instrument	Entry Target	Exit Target
NP1	Right	Top	Front
NP2	Left	Top	Front
NP3	Right	Front	Top
NP4	Left	Front	Top

task. Text instructions appear on top of stereo viewer on the SSC of the dVSim to display the task protocol. The operator must pass the needle 12 times through a set of highlighted entry-exit targets shown on the sponge. There are 4 configurations for the needle passing (NP) as listed in Table 4.13. The user starts the task by picking up the needle from the initial location. The system highlights the driving instrument and entry target. The user drives the needle through the entry target until it pierces and emerges through the exit target. The needle is then, grasped and pulled out through the exit target. Upon pulling out the needle completely, the simulation highlights the next set of driving instrument and entry target or terminates if the 12 NPs are complete.

For this study, one individual labeled the start and end of the constituent maneuvers within the task flow. Table 4.13 shows the combination of driving instrument and entry-exit targets for the different NPs. We have defined the different categories of maneuvers using this information as NP1, NP2, NP3, and NP4. IMS (inter maneuver segments) were also labeled and these include the user's actions in preparation for the next segment.

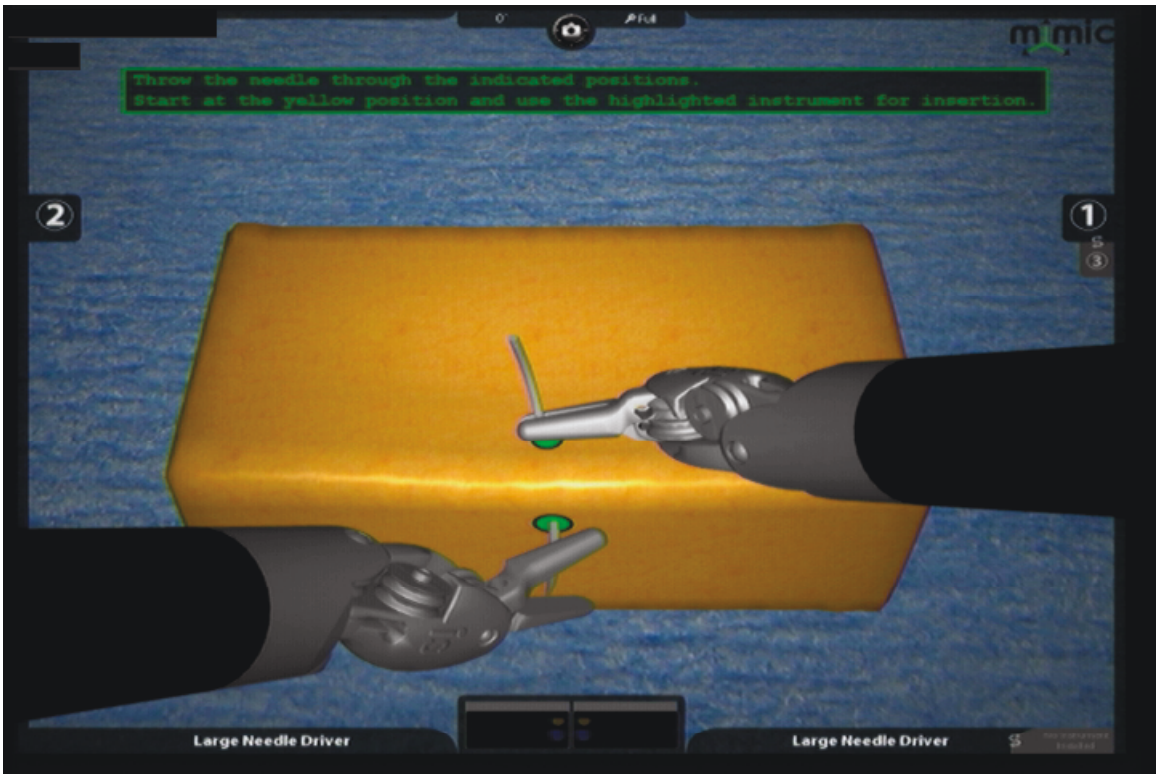


Figure 4.11: A snapshot of the VR needle passing task. Green dots indicate the entry and exit targets for passing the needle.

4.5.2 Preference Annotations

We hosted a web-based HIT (Figure 4.12) to collect preference annotations (A_C) for a set of 100 randomly generated pairs of maneuvers - 20 belonging to each of the five maneuver categories (Table 4.13). We invited members of our research group and colleagues at Johns Hopkins University to participate in the user study via email (similar to the pilot study Section 4.3). Participation was voluntary and there was no reward for participating in the study. The participants were asked to provide a preference indicating which of the two video performances was better, along with their confidence in choosing the preference. This was conducted under an IRB approved study (HIRB00001603).

The HIT layout was the same from the previous validation study in Section 4.4. It was setup to collect at most 5 responses for each pair and was accepting responses for a week. The participants could log on to the web page and work on the HIT for any length of time during that week. Each participant could submit a maximum of one response per pair and potentially submit responses to all the 100 pairs.

4.5.3 Preference Classifier

We chose four subsets of features from Table 4.2 to represent the maneuver performances:

- FeatSet7: CT, 2x PL, 2x RA, 2x MV

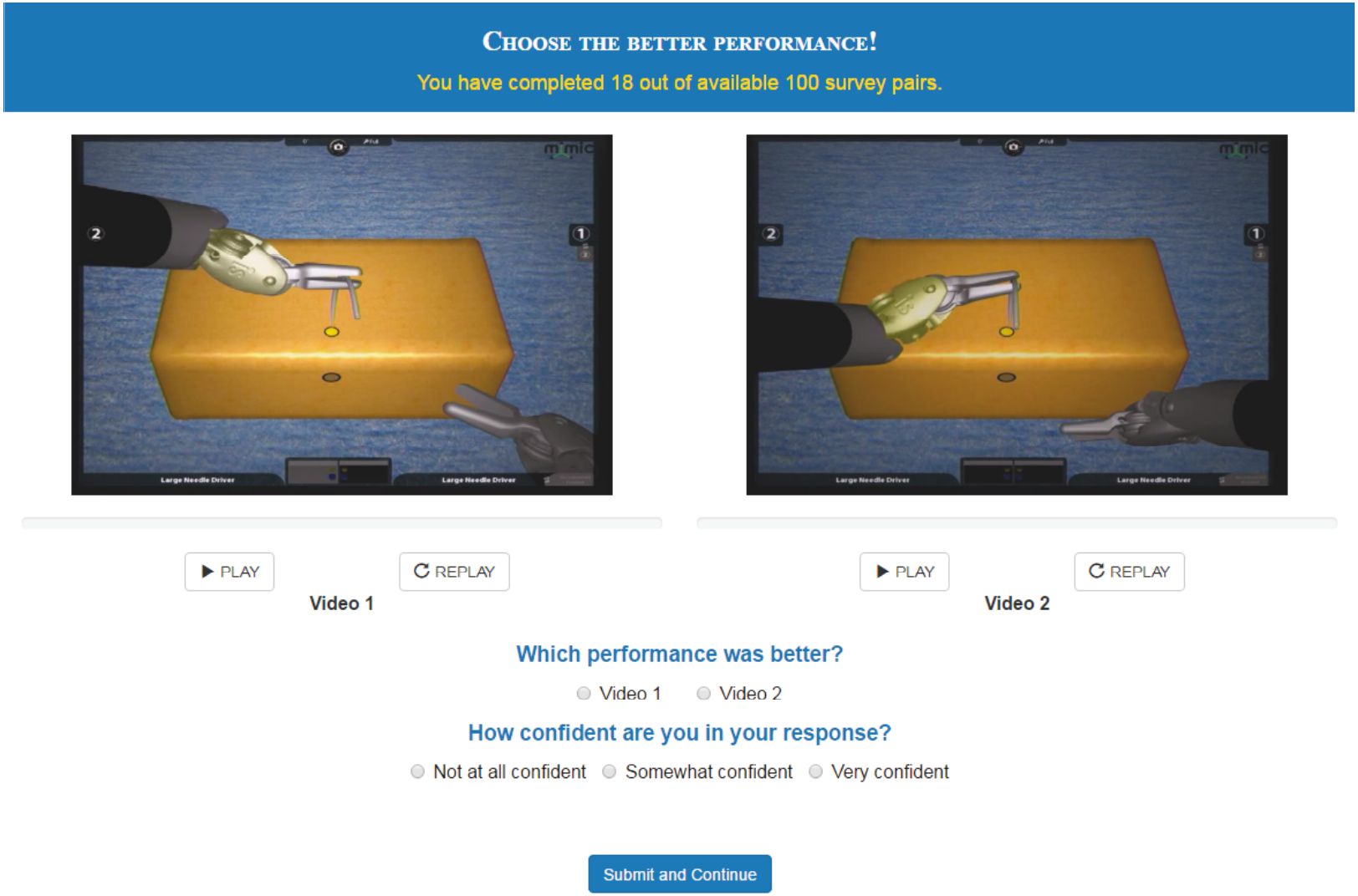


Figure 4.12: A snapshot of the HIT web page

- FeatSet9: FeatSet7 + 2x MPL
- FeatSet11: FeatSet9 + 2x MWV
- FeatSet14: FeatSet11 + TF, 2x GA

A linear SVM was used as before for the preference classifier.

4.5.4 Reliability and Validity Experiments

We measured the inter-rater agreement among the crowd responses using Fleiss' kappa and percent agreement.

We performed two validation experiments – a within data set validation and a cross data set validation. For this we trained preference classifiers using three different training data sources:

1. C_{vr}: using preference annotations A_C obtained in this user study for the VR NP task,
2. C_{bt}: using preference annotations from the previous validation study on bench-top suturing data (Section 4.4), and
3. C_{comb}: using a joint set of preference annotations from VR NP task and bench-top suturing task.

All the experiments were performed in a 5-fold cross-validation setting with 20 iterations. The randomization methods were reset using the same seed to ensure that the test data partitions were same across validation experiments for the different folds

Table 4.14: Inter-rater agreement with 5 raters for each pair of videos

Pairs	N	κ	$perc\ agr$
All	100	0.51 (0.45, 0.57)	85.2 (81.8, 88.6)
$conf \geq 0.5$	88	0.57 (0.50, 0.64)	87.0 (83.5, 90.6)

The values in the parentheses are 95% confidence intervals for the statistic.
κ: Fleiss' kappa, $perc\ agr$: percent agreement

and iterations. A simple measure of accuracy was used as the evaluation metric.

Similar to previous studies, we computed an agreement (agr) and confidence metric ($conf$) for each pair in the survey (refer Section 4.4.3). Samples with $agr \geq 0.75$ and $conf \geq 0.5$ were used for the validation experiments. Additionally, data was scaled to zero mean and unit variance in each experiment using the training data mean and σ (standard deviation).

4.5.5 Results

There was moderate inter-rater agreement in the preference annotations (Table 4.14). The agreement was higher in pairs with high confidence ($conf \geq 0.5$).

Validity of Preference Classifiers

Table 4.15 shows the results from validation of the three preference classifiers C_{vr}, C_{bt} and C_{comb} using the VR NP task as the test data set. C_{vr} had an accuracy of 90% compared to C_{bt} around 88%. The performance of C_{comb} that was trained using both VR and bench-top data was the highest of the three except in case of FeatSet14.

Table 4.15: Preference prediction accuracy on VR NP task of C_{vr}, C_{bt} and C_{comb} using 5-fold cross-validation across different sets of features

Maneuver (N)	Model	FeatSet7	FeatSet9	FeatSet11	FeatSet14
All (62)	C_{vr}	90.81 (0.39)	90.97 (0.46)	93.39 (0.47)	91.69 (0.64)
	C_{bt}	88.06 (0.30)	89.68 (0.34)	86.53 (0.36)	78.06 (0.49)
	C_{comb}	91.85 (0.49)	92.90 (0.47)	94.03 (0.41)	90.89 (0.62)
NP (54)	C_{vr}	91.11 (0.25)	91.02 (0.36)	92.96 (0.42)	91.30 (0.54)
	C_{bt}	91.67 (0.34)	92.04 (0.36)	89.26 (0.39)	80.83 (0.47)
	C_{comb}	92.04 (0.38)	93.43 (0.46)	94.07 (0.35)	90.65 (0.68)
IMS (8)	C_{vr}	88.75 (2.01)	90.63 (2.54)	96.25 (1.31)	94.38 (1.69)
	C_{bt}	63.75 (0.86)	73.75 (2.38)	68.13 (1.43)	59.38 (1.54)
	C_{comb}	90.63 (2.20)	89.38 (2.61)	93.75 (1.70)	92.50 (1.67)

Numbers in parentheses are standard errors over 20 iterations of the cross validation.
In each case, a single preference classifier was trained using available training data.

Relatively lower performance of C_{bt} can be attributed to the fact that it was never trained on IMS in the bench-top data set and can be seen in the last row of Table 4.15.

Results from the cross-validation experiments while testing on the bench-top data set are shown in Table 4.16. C_{bt} has an accuracy of 83% across the different feature sets. The performance of C_{vr} drops from FeatSet7 through FeatSet14. Unlike, the observation about C_{bt} performance on IMS, in this case, C_{vr} performs really well on KT even though it was not trained on any KT sample. C_{comb} performed similar to C_{bt} using FeatSet7 and FeatSet9, but performance dropped drastically using FeatSet11 and FeatSet14. Note, the SE values for C_{comb} in these two cases were larger compared to others. Performance for all the classifiers across the different feature sets showed

Table 4.16: Preference prediction accuracy on bench-top suturing task of C_{vr}, C_{bt} and C_{comb} using 5-fold cross-validation across different sets of features

Maneuver (N)	Model	FeatSet7	FeatSet9	FeatSet11	FeatSet14
All (240)	C_{vr}	80.56 (0.15)	80.54 (0.14)	79.25 (0.13)	76.15 (0.13)
	C_{bt}	83.42 (0.24)	83.15 (0.24)	83.88 (0.26)	83.31 (0.23)
	C_{comb}	83.44 (0.30)	82.50 (0.27)	62.42 (1.40)	65.04 (1.18)
ST (66)	C_{vr}	66.67 (0.25)	63.86 (0.20)	64.09 (0.16)	64.62 (0.28)
	C_{bt}	69.09 (0.69)	69.32 (0.75)	72.42 (0.80)	72.12 (0.63)
	C_{comb}	71.82 (0.76)	70.61 (0.62)	59.70 (1.48)	60.98 (1.41)
GPR (41)	C_{vr}	72.93 (0.53)	78.29 (0.46)	75.61 (0.61)	66.71 (0.44)
	C_{bt}	82.80 (0.21)	82.80 (0.21)	83.05 (0.21)	81.83 (0.33)
	C_{comb}	80.85 (0.44)	81.59 (0.48)	58.90 (2.24)	61.95 (1.87)
KT (133)	C_{vr}	89.81 (0.13)	89.51 (0.20)	87.89 (0.15)	84.77 (0.18)
	C_{bt}	90.71 (0.23)	90.11 (0.25)	89.81 (0.21)	89.32 (0.34)
	C_{comb}	90.00 (0.27)	88.68 (0.32)	64.85 (1.46)	68.01 (1.38)

Numbers in parentheses are standard errors over 20 iterations of the cross validation.
In each case, a single preference classifier was trained using available training data.

the similar trend of lowest performance on ST, followed by GPR and highest on KT.

4.5.6 Outcomes

Our findings in this study establish validity of our framework for segment-level technical skill assessment within VR, bench-top and across bench-top and VR platforms. The high preference prediction accuracy observed with C_{bt} and C_{comb} in Table 4.15 and with C_{vr} and C_{comb} in Table 4.16 indicates that the metrics we used and pairwise comparisons of segments capture aspects of technical skill that were consistent across the bench-top and VR platforms. Subsequent research should assess whether

transferability of skill assessment metrics that we observed in this study extends to operating room data.

4.6 Absolute v/s Relative Ratings: Reliability Analysis

Over the course of this chapter, we presented three validation studies for our proposed framework to generate automated and objective skill assessment scores at segment- and task-level. Our scoring framework relies on the core concept of pairwise comparisons to rank a given performance among previous performances present in a corpus $\mathcal{L}$ (Section 4.2.2). We adopted this pairwise comparison strategy due to three reasons – (1) there is no existing standard for skill assessment at segment-level, (2) obtaining ground truth using experts (surgeons) is not scalable, and (3) intuitively, crowd (not necessarily surgically trained) should be better at relative rating compared to absolute rating. We have shown comparable inter-rater reliability between crowd and experts as well as validity of pooled crowd responses compared to pooled expert responses on relative ratings. However, we assumed that relative rating is easier (more reliable) than absolute rating. In our knowledge, there are no previous works or evidences to support this assumption. With this goal in mind, we designed a crowdsourcing study to collect absolute and relative ratings and test their reliability.

4.6.1 Data Set

We used the FESS Targeting data set described in Appendix A.4 for this study. This task was different in two aspects compared to the previous studies: (1) this was a camera targeting and pointing task compared to needle manipulation tasks, and (2) a know-how of anatomy was required to evaluate surgical skill compared to broadly technical skills previously.

Video segments for different targeting sub-tasks were extracted resulting in a data set of 371 targeting performance videos across the nine targets (Table A.3). An individual watched all the videos and annotated segments wherein the endoscope was out of the cadaver head. All such video frames were replaced with a black frame showing the text "`CAMERA OUT OF PATIENT HEAD`". This was done to ensure anonymity and be compliant with the IRB committee's guidelines and recommendations.

4.6.2 Crowdsourcing Study

In this study, we used the MTurk (Section 2.4) platform to recruit annotators for both absolute and relative ratings. In the previous studies, performance videos contained technical skills training tasks on inanimate objects. Skill assessment in such scenarios does not require medical / anatomical knowledge. In contrast, the FESS targeting task was designed to test anatomical knowledge about the target's location and technical skills to navigate inside the sinus cavity using the endoscope

and nasal pointer instrument. Thus, we decided to train and orient the workers about the task and anatomy before collecting skill ratings from them. Like the context summarization study (Section 3.6, Figure C.2), we followed the training ⇒ testing ⇒ before collecting skill responses protocol.

Custom Qualifications

We created nine custom qualifications (CQ), one per target. Let us describe the qualification process for Target 1. A qualification HIT (QH_1) was created on MTurk with a reward of \$0.1, duration of 60 minutes, qualification requirement (AdultContent = 1, TotalHITsApproved $\geq$ 100, HITApprovalRate $\geq$ 95, refer Section C.3). The QH_1 web page showed a training video with narration describing the FESS task and specifically the Target 1 showing examples of a good and bad execution. The web page also displayed the informed consent, study description and instructions about the HIT. The workers were required to watch the training video from start to end at least once, before they could take the test. The training video link was provided during the test for quick reference. The test included a performance video along with the questions that appear on the absolute rating survey (Figure 4.13). The workers were given two attempts to answer correctly. They were prompted about one of the answers being wrong without specifying the particular question. Upon passing the test, the workers were granted CQ_1 and a bonus payment of \$0.3 for the time spent on the QH_1. Upon failure, the workers were granted a block qualification to pre-

vent them from re-attempting QH_1. They were still allowed to attempt other target qualifications.

Absolute Ratings

We setup HITs on MTurk to collect responses using an absolute rating (ABS[5]) approach on all targeting performances in our data set. Multiple HITs were set up and each HIT was specific to a target with the CQ_i required for target i. A sample survey form is shown in Figure 4.13. A performance video along with a reference target image is shown on top. The workers are asked to provide binary skill ratings across four components – accuracy, smoothness (in motion), tissue awareness and efficiency (in motion) and an overall binary rating as well. As shown in the sample form, the workers were asked to mark their confidence levels on these ratings as well. The workers could submit their ratings only if they had finished watching the video completely. Depending on the number of videos setup in the HIT, they would move to the next video evaluation until HIT completion. Attention videos were pre-selected and appeared at random to check for spam from workers. The worker's responses on such attention videos were checked against an answer key. A failure on the attention video resulted in termination of the HIT along with a revocation of CQ_i. Training video link (not visible in the image) was available on the top of the page as well.

[5]we will use abs/ABS as a notation for any reference to absolute ratings in future

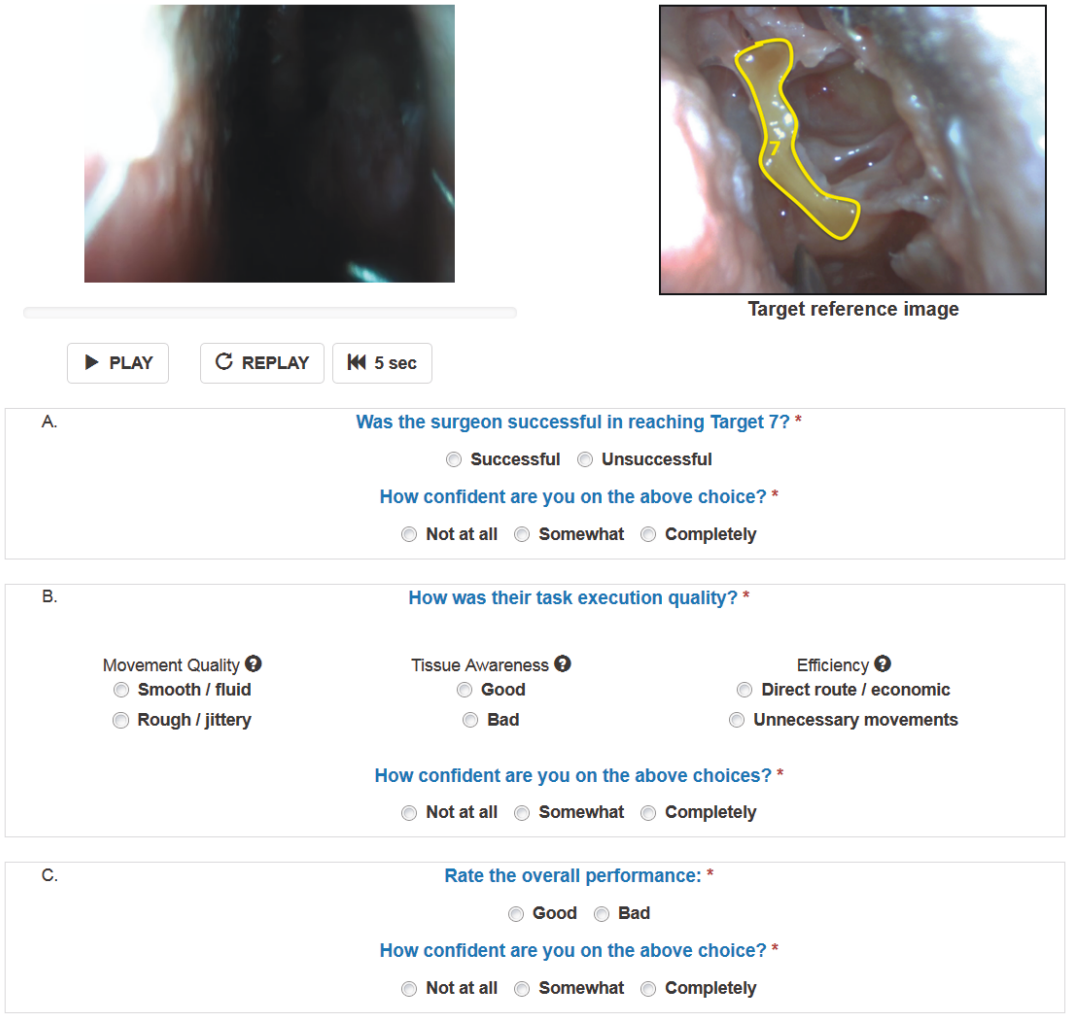


Figure 4.13: A snapshot of the absolute rating HIT questionnaire

Relative Ratings

In case of relative ratings, we chose a limited number of pairs from the data set. We selected 10 training sessions (out of 49) by sampling $3-2-2-3$ sessions at random belonging to the expert-assigned skill rating intervals $[1,2)-[2,3)-[3,4)-[4,5]$ respectively (refer to Appendix A.4). All possible pairwise comparisons of matched target performances were extracted (a matched pair would contain a comparison of target 1 versus target 1 and not any other target) resulting in a total set of 326 pairs across the nine targets.

Multiple HITs were set up and each HIT was specific to a target with the CQ_i required for target i. A sample survey form is shown in Figure 4.14. Two performance videos of Target i are shown on top along with a reference target anatomy image. The workers provided their preference along with a confidence level. Depending on the number of pairs per HIT, they would move to the next pair until completion. Attention pairs were pre-selected and appeared at random to check for spam from workers. The worker's responses on such attention pairs were checked against an answer key. A failure on the attention pairs resulted in termination of the HIT along with a revocation of CQ_i. Training video link (not visible in the image) was available on the top of the page as well.

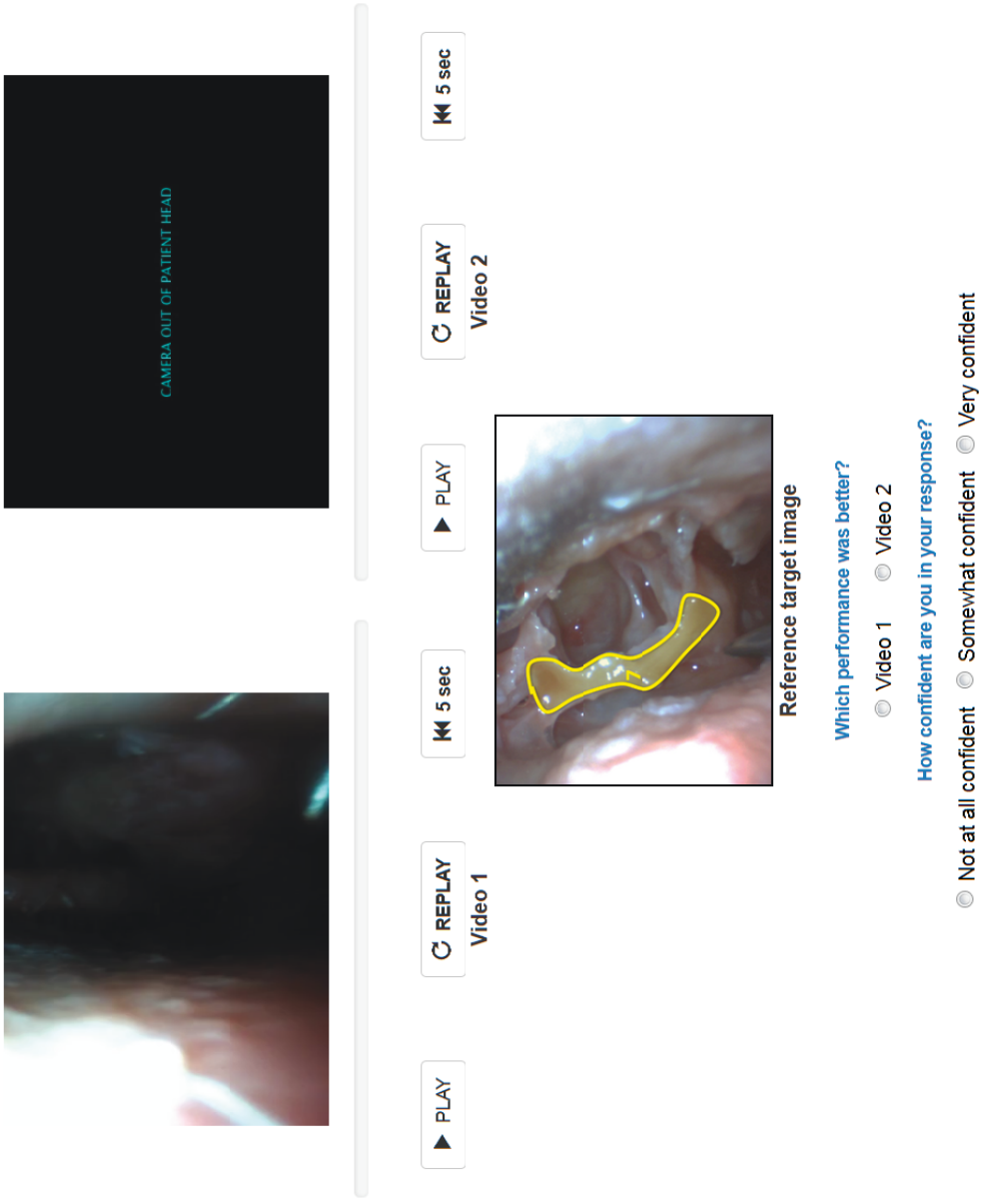


Figure 4.14: A snapshot of the relative rating HIT questionnaire

Pilot Study

As mentioned earlier, the FESS training task involves assessment based on correctness of anatomy as well as technical skills. Since none of our previous studies involved correctness of anatomy, we conducted a pilot study to estimate the number of workers required for achieving high inter-rater agreement and validity. A subset of 36 tasks were chosen at random (four belonging to each target) for absolute ratings (from the pool of 371 videos) and for relative ratings (from the pool of 326 pairs). A total of 25 responses were collected on each of the 36 absolute and relative rating tasks.

We computed inter-rater agreement using percent agreement and Fleiss' kappa. We compared absolute ratings for the overall question against previously collected expert ratings for measuring validity. Three experts had provided binary absolute ratings on the overall question for 32 of these 36 pilot tasks. There were no expert assigned relative ratings for measuring validity. To estimate the reliability and validity of crowd ratings for smaller number of workers, we sub-sampled the responses (without replacement) and computed the inter-rater agreement (absolute and relative) and validity (absolute - overall question). We repeated the sub-sampling over 20 iterations to obtain a confidence interval.

Table 4.17: Inter-rater agreement observed in the pilot study

Rating Method	Question	*perc agr*	κ
Absolute	Overall	84.14 (2.34)	0.48 (0.01)
	Accuracy	84.48 (2.35)	0.44 (0.01)
	Smoothness	84.00 (2.35)	0.48 (0.01)
	Tissue Awareness	83.05 (2.47)	0.34 (0.01)
	Efficiency	85.30 (2.33)	0.50 (0.01)
Relative	Preference	80.86 (2.76)	0.40 (0.01)

values in the parentheses are standard errors

Results

Table 4.17 shows the inter-rater agreement on the absolute and relative ratings among the crowd workers. Inter-rater agreement among the expert ratings was 88.54 (SE: 2.95) using percent agreement and 0.47 (SE: 0.10) using Fleiss' kappa.

We observed a lower agreement on relative ratings compared to absolute ratings. We investigated the underlying confidence level (refer Section 4.4.3) for these tasks and computed the inter-rater agreement among responses for tasks that were marked with a certain minimum confidence level. Figure 4.15 shows the outcome of this analysis. We saw that a larger proportion of relative tasks were marked with lower confidence. The agreement for relative ratings and absolute ratings are similar if we compare responses from higher confidence marked samples.

Validity of the crowd responses for the overall question in absolute ratings was 81.25% compared against expert ratings on 32 tasks.

Table 4.18 shows that the inter-rater agreement stays consistent while varying the

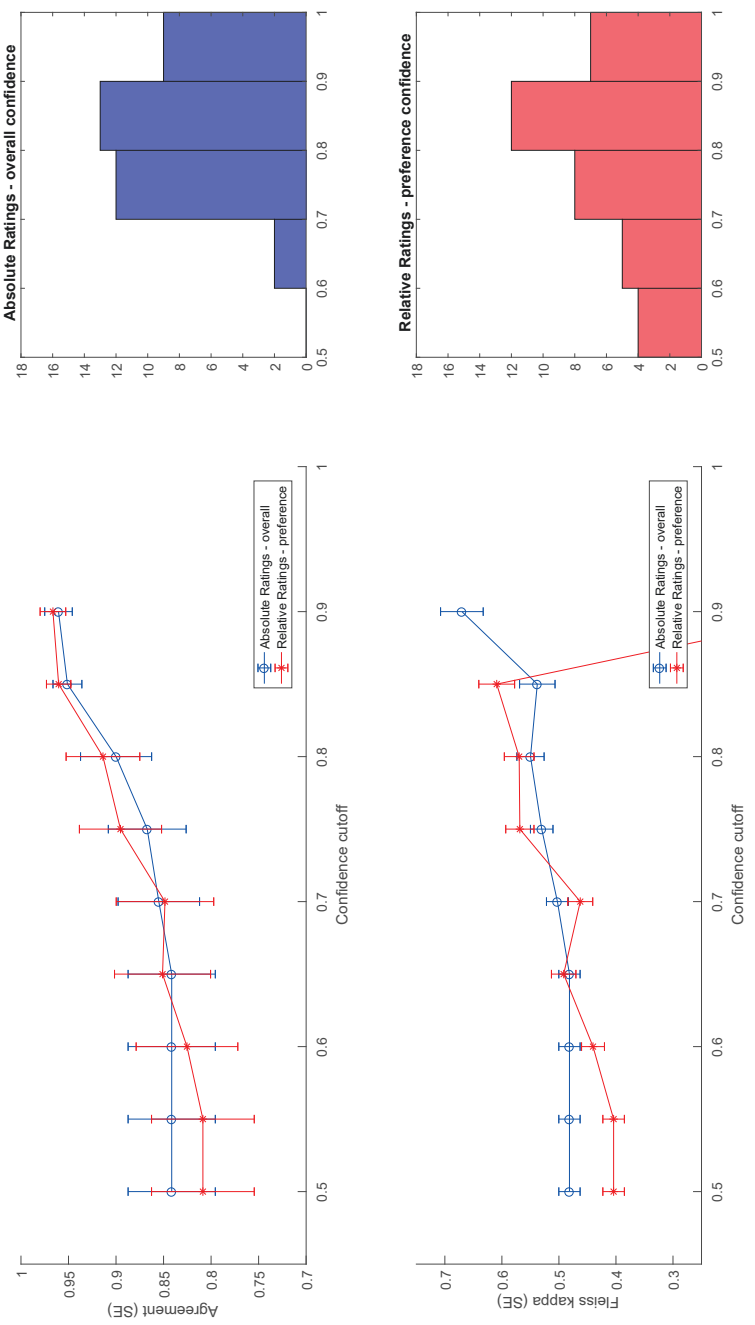


Figure 4.15: Comparison of inter-rater agreement for absolute and relative ratings. Agreement values are computed using tasks that were marked above the confidence level on x-axis. The bar plots show the distribution of confidence level across the tasks.

Table 4.18: Inter-rater agreement v/s number of workers using Fleiss' kappa

		# of workers N			
Rating Method	Question	5	11	17	23
Absolute	Overall	0.47	0.49	0.48	0.48
	Accuracy	0.43	0.44	0.45	0.44
	Smoothness	0.49	0.48	0.48	0.48
	Tissue Awareness	0.31	0.35	0.34	0.34
	Efficiency	0.49	0.50	0.50	0.50
Relative	Preference	0.38	0.39	0.41	0.40

Table 4.19: Validity of crowd responses v/s number of workers on the overall question for absolute ratings

# of workers (N)	5	11	17	23
Accuracy	79.06	81.56	80.94	80.16

number of workers across the absolute rating questions and relative rating preferences.

Table 4.19 suggests that the validity of crowd responses increases as we gather responses from more crowd workers (N) initially and plateaus after $N = 11$.

Figure 4.16 shows the inter-rater agreement and validity for the absolute rating overall question for different values of N (number of workers). We can observe that the agreement and validity start plateauing around $N = 9$.

4.6.3 Reliability Study

Based on the pilot study results, we collected responses from a maximum of 9 workers on each of the remaining 335 absolute and 290 relative tasks. We chose an

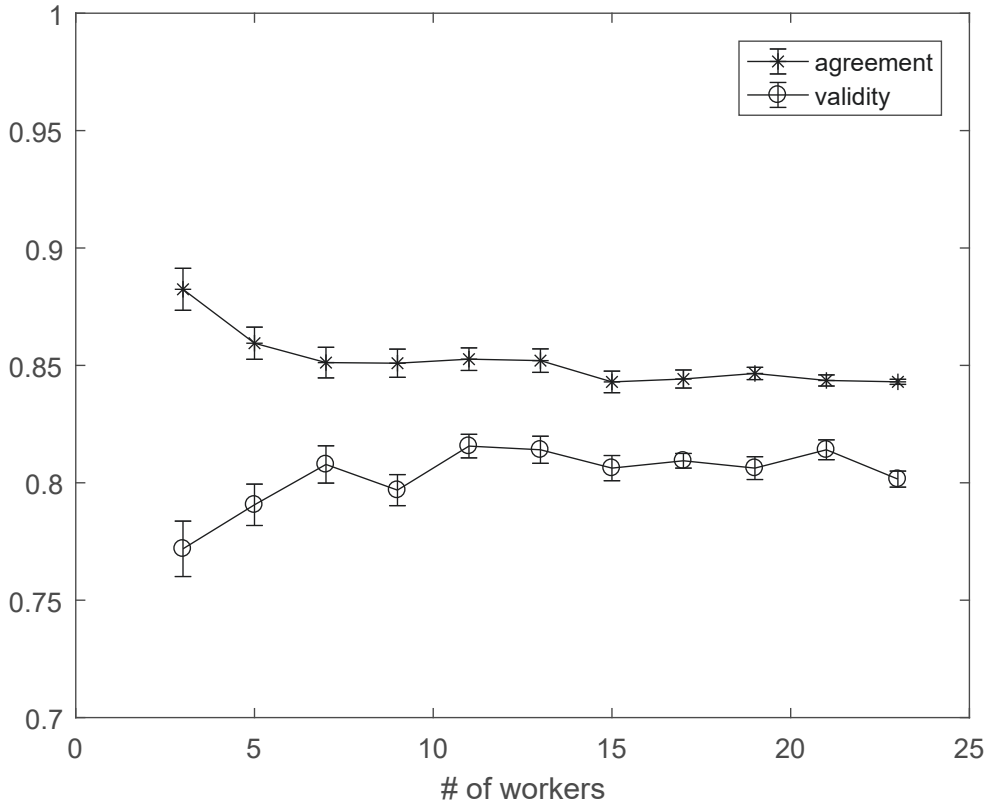


Figure 4.16: Inter-rater agreement and validity for overall question of absolute ratings v/s number of workers

adaptive strategy to collect responses starting from three workers. The HIT web page was setup to collect a minimum of three responses per task. The percent agreement (agr) for the task using the responses from the three workers was computed. If $agr \geq 0.75$, then the framework stopped collecting any more responses on that task. If $agr < 0.75$, then it collected two more responses. This process continued until $agr \geq 0.75$ or responses from nine crowd workers had been collected.

We measured the inter-rater agreement for the absolute and relative ratings obtained using percent agreement and Fleiss' kappa. We also measured inter-rater agreement with different confidence level cutoffs i.e. computing inter-rater agreement over tasks with a minimum level of confidence assigned by crowd ratings. We were able to compute the validity of the absolute ratings for tasks that had been previously rated by three experts.

Results

Table 4.20 shows the inter-rater agreement for absolute and relative rating tasks. We observe that the inter-rater reliability between absolute and relative ratings is quite similar.

Figure 4.17 shows the inter-rater agreement for absolute and relative tasks based on the confidence levels assigned to the task by crowd raters. We see that relative ratings are more reliable than absolute ratings with non-overlapping confidence intervals for higher confidence level tasks.

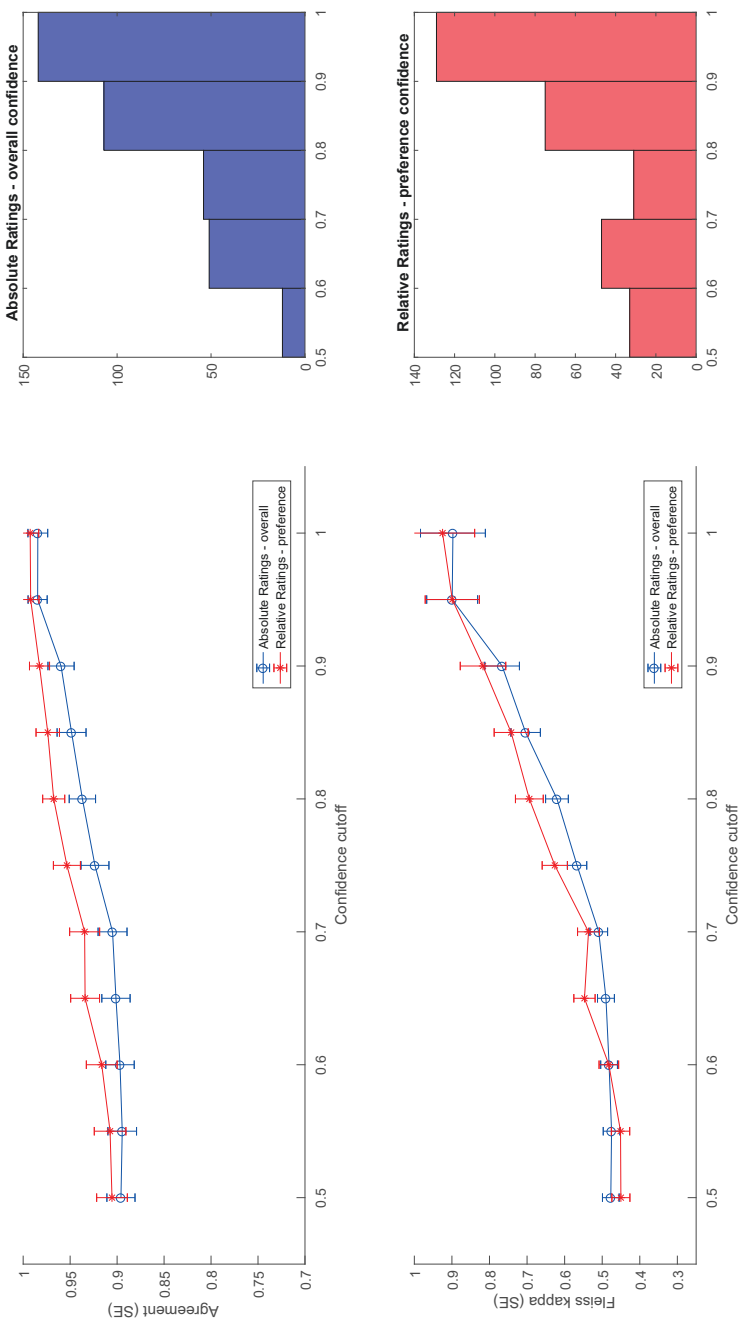


Figure 4.17: Comparison of inter-rater agreement for absolute and relative ratings. Agreement values are computed using tasks that were marked above the confidence level on x-axis. The bar plots show the distribution of confidence level across the tasks.

Table 4.20: Inter-rater agreement observed in the main study

Rating Method	# of tasks	Question	*perc agr*	κ
Absolute	371	Overall	89.67 (0.76)	0.48 (0.01)
		Accuracy	89.61 (0.76)	0.47 (0.01)
		Smoothness	84.72 (0.86)	0.46 (0.01)
		Tissue Awareness	84.03 (0.87)	0.33 (0.01)
		Efficiency	86.38 (0.82)	0.49 (0.01)
Relative	326	Preference	90.04 (0.83)	0.44 (0.01)

values in the parentheses are standard errors

The validity of the crowdsourced absolute skill ratings against the pooled expert-assigned rating was 79.93% (95% CI: 75.41 to 84.46) on a subset of 304 tasks that had been rated by both the experts and the crowd. In comparison, the experts' inter-rater agreement was 90.46% (95% CI: 88.75 to 92.17).

4.6.4 Outcomes and Future Work

In this study, we compared absolute and relative ratings for the FESS targeting task. We observed moderate inter-rater agreement for both absolute and relative ratings tasks. Filtering out low confidence level annotations, relative ratings had higher inter-rater agreement compared to absolute ratings.

Inter-rater measurements are one component of reliability analysis. Intra-rater reliability is equally important and future work should focus on comparing absolute and relative ratings on intra- and inter-rater agreement. However, it should be noted that intra-rater reliability in the setting of crowdsourcing can be a bit ambiguous.

Typically, crowdsourced annotations are pooled from multiple raters. Intra-rater and inter-rater reliability in case of the crowd as a single (pooled) rater collapse into a single measurement. Intra-rater reliability can be measured for the individual members of the crowd.

Additionally, a comparison of the validity of absolute and relative ratings should be performed against expert-assigned ratings for a complete comparison of the two rating methods.

Absolute ratings in our study were performed using binary ratings. However, other assessment tools like OSATS and GEARS (which are current standards for global assessment) use Likert-like scales. Future work should collect Likert-like scale-base absolute ratings and compare them against pairwise comparisons-based relative ratings.

As the goal of collecting such ratings on segment-level tasks is to train machine learning algorithms, validity comparison of such models should be performed using absolute and relative ratings as training data.

4.7 Discussion

Our framework currently relies on pre-existing surgical task segmentation into constituent maneuvers. While previous research on recognition of surgical activity exists, very few researchers have looked at maneuver level activity recognition.[126,127]

Our work illustrates the need for reliable and accurate technology for automatic recognition of surgical maneuvers. Future work should look at the accuracy required from such methods to perform valid and useful segment-level evaluation using our framework.

Manually assigned GRS reflects a global assessment of task performance by an experienced surgeon and includes elements such as knowledge of task, respect for tissue, and forward planning. The objective skill scores generated using our framework did not account for all items considered when a surgeon manually assigns a GRS to a trial because of the few elements in the feature vector that we used to train the preference classifier. Future work combining different data modes and other metrics should be performed to capture skill information from the other components of the GRS.

Our work leads to additional future research. A library of maneuver performances ($\mathcal{L}$) is required for our framework to generate scores that provide relevant and meaningful feedback to trainees. What are the characteristics – composition, size – of this library? Do we need a different library for different tasks or skills? These are all questions that need to be answered.

One very interesting open question is whether pairwise comparisons provide a more effective means for crowdsourced skill assessment than global assessments, and whether the effectiveness of the framework is sensitive to the granularity of analysis. Conversely, the most effective level of analysis for teaching is also not yet established.

Feedback at levels finer than maneuvers in the task, such as gestures, may be important for surgical skills acquisition. For example, errors in performance of the task are typically articulated at the gesture-level, and thus, gesture-level assessments using our framework may yield effective feedback for trainees. The effectiveness or educational value of gesture, maneuver, and task-level assessment for acquisition, maintenance, and retention of surgical technical skills remains to be investigated in future studies.

Finally, we note that our approach may be deployed on any surgical platform where we can capture the data necessary to compute quantitative measures of surgical skill. This includes robotic, open, conventional laparoscopic, and endoscopic surgery. For example, Ahmidi et al.[165] capture motion data in an open procedure to preform reliable skill assessment.

4.8 Summary

In this chapter, we presented a background on current state of surgical skill assessment and the shortcomings related to segment-level assessment and directed feedback (*where was I wrong?*). We outlined the challenges in segment-level evaluation and motivated the need for a solution. Towards that end, we presented a framework for crowdsourced skill assessment that yields valid objective surgical skill assessments both for the overall task and for maneuvers within a task.

Previous works on automated surgical skill assessment have focused on global

(task) level performance. Few have shown success in predicting skill scores while the rest have been limited to skill class predictions (novice, intermediate and expert). We have proposed and validated a skill scoring framework for task and segment level performance which is a first in existing literature. Likewise, crowdsourcing for surgical skill assessment was limited to global evaluation using absolute rating tools. We presented a relative ratings approach for segment level evaluation and validated its accuracy and reliability for the first time.

Specifically, we showed that crowdsourcing can provide reliable pairwise comparisons for maneuvers within a task, and that pairwise comparisons by a surgically untrained crowd used within our framework yield segment- and task-level assessments that are comparable to those obtained using pairwise comparisons by expert surgeons. We showed that preference classifier learned using preference annotations from a bench-top suturing task provides valid comparisons of performances in a VR needle passing task and vice versa. We showed that the reliability and validity of findings obtained from applying our framework hold across the size, member constitution, or other properties of the crowd as well as surgical task. Finally, we showed preliminary results indicating higher inter-rater agreement in assigning relative ratings compared to absolute ratings using a FESS targeting task data set.

We have presented a valid and reliable solution to generate segment and global skill scores to deliver the coaching intervention of – *how did I do it?* and *where was I wrong?*

Chapter 5

Feedback and Teaching

Table 5.1: Individualized feedback and context-relevant demonstration are important for teaching mastery of skill

Training Type	Activity Recognition	Skill Assessment	Feedback	Demonstration
VR	✔	🎓[4,5]	🎓[5]	🎓[5]
Bench Top	✔	🎓[4]	✖	✔
OR	✔ 🎓[3]	✔	✖	✖

✔: indicates previous work in literature exists, 🎓[X]: indicates a solution is presented in Chapter "X" of this thesis, ✖: indicates no prior work exists and none is presented in this thesis.

To a trainee, the realization that something was wrong in their performance and identification about where they were wrong is quite important (and we looked at addressing these problems in Chapter 4). However, the understanding of why were they wrong is the next most important piece of information that they want to know. Without knowing the reason behind their sub par performance, the process of learning

and improvisation is halted. This ties to the next activity of a coach listed in Section 1.3 – the ability to provide an answer for *why was I wrong?*

This question of *why*, is in fact, quite closely tied to another question – *how do I do it correctly?* Along with knowing why the performance was poor or wrong, an insight to what is the correct approach is ideal for effective learning. In other words, it is important from the perspective of the trainee to observe proficiency to strive towards being proficient. **Demonstration (*how*) and feedback (*why*) go hand in hand. A feedback without any demonstration seems incomplete, while a demonstration without any feedback seems irrelevant or meaningless.**

In this chapter, we will present our approach to provide feedback and demonstration (teaching) in the setting of a virtual reality (VR) simulation-based training. But first, let us motivate the need for feedback and teaching in surgical training.

5.1 Background

Similar to the need for objective skill assessment tools, there has been advocacy of feedback-based training and assessment for the development and testing of surgical proficiency. Feedback can be presented in various forms and by multiple agents.

Firstly, feedback can be "internal" also referred to as "self-proctored" or "self-assessed" or "independent". Basically, the trainee is left in an independent learning environment without any instructor or mentor overlooking their performance. In the

true sense of the concept of 'internal feedback', the trainee should be measuring their performance and learning by themselves. However, with development of computer technology and simulation environments, a corrupted version of the concept includes the use of scores generated by the computer-assisted learning systems. The trainee still comprehends the scores, detects inefficiencies or errors and lays out learning goals by themselves.

The other type of feedback is "external", wherein the agent is an instructor who reviews and provides feedback to help improve the trainee's performance. External feedback can be expert or non-expert based and can be delivered concurrently, immediately or delayed.

Likewise, various forms of feedback are available – verbal or audio-based, visual (video-based), haptic (force-based) or check-list based.

The value of feedback has been studied and advocated in other domains of health care as well. Wigton et al.[182] conducted an experiment with medical students while teaching diagnosis using simulated patient data. The group of students that was provided post-diagnosis feedback about the correct weighting of information showed improved judgment and accuracy in diagnosis compared to the control group. Salas et al.[183] listed 'providing feedback' as a key factor for the success and effectiveness of simulation-based health care team training for improving patient safety.

5.1.1 Feedback in Surgical Training

A large number of studies have been conducted to study the effectiveness of different types of feedback mentioned above.

Rogers et al.[184] showed that external feedback by a content expert is better for performance learning than computer-assisted learning alone. They noted that the output of the study task (knot tying) was comparable between the groups, however, independent learners were not able to identify the errors and outline steps necessary to improve performance. In case of Mahmood et al.,[185] no improvement in performance on a colonoscopy simulator was observed in absence of feedback.

External constructive feedback by an expert surgeon showed better performance improvement when compared to no feedback in the works of Grantcharov et al.[66] (laparoscopic cholecystectomy in OR), Hamad et al.[67] (laparoscopic jejunojejunal anastomosis in OR), Porte et al.[68] (knot tying and suturing skills in simulation), Kruglikova et al.[69] (colonoscopy simulation), Boyle et al.[70] (hand-assisted laparoscopic colectomy in simulation), Boyle et al.[186] (renal artery angioplasty in simulation), Strandbygaard et al.[71,187] (laparoscopic salpingectomy in simulation), Ahlborg et al.[72] (laparoscopic surgery in simulation) and Vaughn et al.[73] (knot tying and suturing skill in simulation).

Specifically, concurrent feedback was studied by Ahlborg et al.[72] All of the other works were based on post-completion feedback. Stefanidis et al.[188] compared limited (10-15 minutes) versus intense (1-2 hours) feedback and showed that the former was

superior to the latter. Boyle et al.[186] compared no feedback, expert-based feedback and non-expert feedback, and concluded that nonexpert facilitators can also enhance the quality of training and may represent a valuable alternative to expert clinical faculty. Vaughn et al.[73] compared performance of interns based on whether they received feedback from their peers (PF) or from faculty (FF). The PF group performed better at the final assessment, suggesting a potential advantage of skill development through reviewing and critiquing others' performance. In addition, Bjerrum et al.[189] and Porte et al.[68] showed expert feedback-based training had better skill retention compared to no feedback or independent learning. While, Ahlborg et al.[72] studied a joint concurrent and immediate feedback group versus no feedback. One of the studies[190] found that proctored training had no advantage over independent learning.

A common goal of all the above works was to demonstrate the value of feedback in surgical proficiency acquisition and testing. Some of the works recommended the development of stand-alone simulation systems to provide feedback and demonstration. Porte et al.[68] suggested that the availability of pre-recorded expert demonstrations and motion efficiency feedback, along with appropriate instruction in self evaluation will prove to be a more complete and effective educational method. Kruglikova et al.[69] proposed that future software developments should aim at providing improved constructive, real-time feedback thus eliminating the need for a supervising expert at each teaching session.

Automated Feedback Mechanisms

To the best of our knowledge there are no existing automated methods that deliver expert-like feedback for surgical training. Though, various mechanisms to provide indirect feedback have been explored and validated in research.

Reiley et al.[74] studied the effects of visual force feedback, a haptic feedback surrogate, on tying surgical knots. Their system measured force applied by the instrument on the task model and displayed a visual scale on the monitor to indicate excessive force application to the surgeon. They showed that such visual force feedback resulted in reduced suture breakage, lower forces, and decreased force inconsistencies among novice robotic surgeons. No advantage was observed among experienced surgeons.

Virtual fixtures (VF)[191] are another mechanism to provide active feedback in the form of constraining motion of the surgical tool on a desired path or within a desired region. The concept was developed as a robot-assistance tool; however, it has shown value in skill learning. Chen et al.[15] demonstrated the use of VF for development of surgical skills in a robot-assisted suturing and knot tying simulation task. Vedula et al.[192] compared learning specific tasks in sinus surgery with and without VF.

While these mechanisms have been applied in surgical training activities, a big limitation is that they are currently static and non-individualized. Further development of these concepts and validation for surgical training and mentoring is a possible future research direction.

In conclusion, previous works have proved value of expert-based feed-

back in learning and recommended a need for automation of constructive and real-time feedback, while there are still no existing tools that deliver such feedback for surgical training.

5.1.2 Error Analysis

To err is human

Identifying errors in performance is inherent to delivering effective feedback. Error spotting and detection are part of the answer to the question *why was I wrong?*

It has been previously reported[193] that medical errors are the eighth leading cause for deaths in the United States with approximately 98,000 preventable deaths per year. Prior works have shown identifying, explaining and demonstrating errors is of great learning value for surgical residents.[34,194] DaRosa and Pugh[33] justified the importance of integrating planned instruction about errors into surgical residency curricula despite reduced work hours and an already overcrowded residency curricula. Fischer et al.[195] surveyed residents and medical students and found that none of them said that they learned better from near misses than from actual errors. In fact, many trainees believed that they learned the most when harm was caused. The authors suggested that future multi-institutional work should focus on learning from errors and near misses. King et al.[196] proposed a new approach for delivering team

training by encouraging errors in low-risk settings like simulation compared to live patients. Rogers et al.[49] demonstrated that instruction about common errors, when combined with instruction about the correct performance enhanced the acquisition of this surgical skill.

We believe that a system to deliver automated effective feedback should be able to measure errors in performance and present the findings along with remedial techniques to reduce such errors. For this, we need to develop error-based as well as context-based metrics for assessment of surgical proficiency. Current VR simulation frameworks do report errors in performance. Though, they are presented more as a score and less as a feedback mechanism.

Detection of errors, however, is a challenging task. While automated methods to perform surgical activity modeling and surgical skill assessment have been developed and are an active area of research in computer assisted interventions, no known work has focused on automated detection of errors in task performance. This goes back to the notion of information contained in different data modalities (Table 3.2). In case of error detection, it is a pre-requisite to have context information. Thus, video-based methods should be developed for this purpose. However, current computer vision methods in surgical data science have operated at the temporal level to model surgical activity with very limited works on surgical context extraction. For this reason, we have developed our methods on feedback, teaching (demonstration), error-based metrics and deficit-based metrics in the virtual reality (VR) simulation setting.

5.2 Framework

5.2.1 Learning Elements

We believe that every skill constitutes of learning elements that require proficiency to perform the skill. The following are characteristics of such learning elements:

- They are **elementary**. For example, length of suture tail and grasp position on the suture to tie the knot are elementary concepts in the skill of suturing.
- They are **consequential**, that is, performing any such element with sub par proficiency can lead to an overall sub par performance. If suture tail length is too short, extra manipulation, time and may be unnecessary force on suture is needed to successfully complete the throw. In the worst case, a re-attempt is required if the suture tail slips through the tissue.
- They are **skill specific**, however some are shared across skills. For example, learning the right length of suture tail is not critical for a successful dissection. However, grasping objects in an ideal manner is critical in different skills like suturing, dissection, sealing and clipping.

5.2.2 Error and Deficit Metrics

As motivated previously, analyzing errors in task performance is key to feedback. The learning elements (Section 5.2.1) in a skill can be associated with failures (errors)

and deviations (deficits). We introduce the concept of error metrics and deficit metrics here.

Error Metrics

As the name suggests, these metrics are defined as counts of errors in task performance. They may be associated with failures in performance of learning elements. Likewise, errors can be generic across skills and specific to certain skills. For example, dropping an object is a generic error, but bleeding due to wrong dissection is an error specific to the skill of cautery usage.

Deficit Metrics

We define *deficit metrics* as measures of deviation from ideal performance of learning elements (Section 5.2.1). For example, ideal behavior in a learning element of grasping may be defined as being normal to the object being grasped. The associated deficit metric to grasping will be the deviation of grasp angle from the normal direction.

Needle Passing Metrics

The American College of Surgeons (ACS) and Association for Program Directors in Surgery (APDS) have jointly developed a set of skills under the "ACS/APDS Surgery Resident Skills Curriculum" (SRSC). Different suturing techniques are part

of Phase I (Core Skills) modules. Each module contains expert demonstrations, step-by-step instructions, evaluation criteria and rating scales.

We have defined error and deficit metrics for needle passing in line with recommendations from Modules 3 (Suturing), 13 (Advanced Laparoscopy Skills) and 14 (Hand-sewn Bowel Anastomosis). We consider a simple needle passing training task wherein the trainee is provided needle driver instruments, a needle and marked entry-exit target pairs on a simulated tissue surface.

I. Error Metrics

Drops: This is the number of times the needle is dropped by the trainee.

Incorrect Targets: VR training tasks typically have an underlying task protocol. The task may require the trainee to perform needle pass through the target pairs in a specific order as in the ISI-SG-Sim Needle Passing data set (Section A.2). This metric captures the number of times the trainee passes needle through the wrong pair of targets.

Missed Targets: This is the number of times the needle is pierced through the tissue outside of the target radius.

Excessive Needle Pierces: This is computed as total number of times the needle pierces the tissue minus 2× number of needle passes to be performed.

Excessive Insert Needle Pierces: This is computed as total number of times the needle pierces the tissue while being inserted in entry target minus number of needle

passes to be performed.

Excessive Exit Needle Pierces: This is computed as total number of times the needle pierces the tissue while being pulled out from exit target minus number of needle passes to be performed.

Excessive Instrument Force: Although this metric is not directly an error, it relates to the error of suture breakage. There are two values associated to this metric - a count and a duration.

- The count is incremented every time force applied by an instrument on an object like needle, other instrument or tissue, crosses a preset threshold.
- The duration is the sum of time intervals during which force applied by an instrument stays above the threshold.

A separate value for this metric is calculated for each instrument in the task.

Excessive Needle-Tissue Force: Although this metric is not directly an error, it relates to the error of tissue tear. Like above, there are two values associated to this metric - a count and a duration. They are incremented based on the force applied by the needle on the tissue.

II. Deficit Metrics

Grasp Position: This is the mean of deviations from the ideal grasp position of the needle across all needle passing attempts. Ideal grasp position varies across different suturing styles and the training task setup. In any case, grasping the needle too

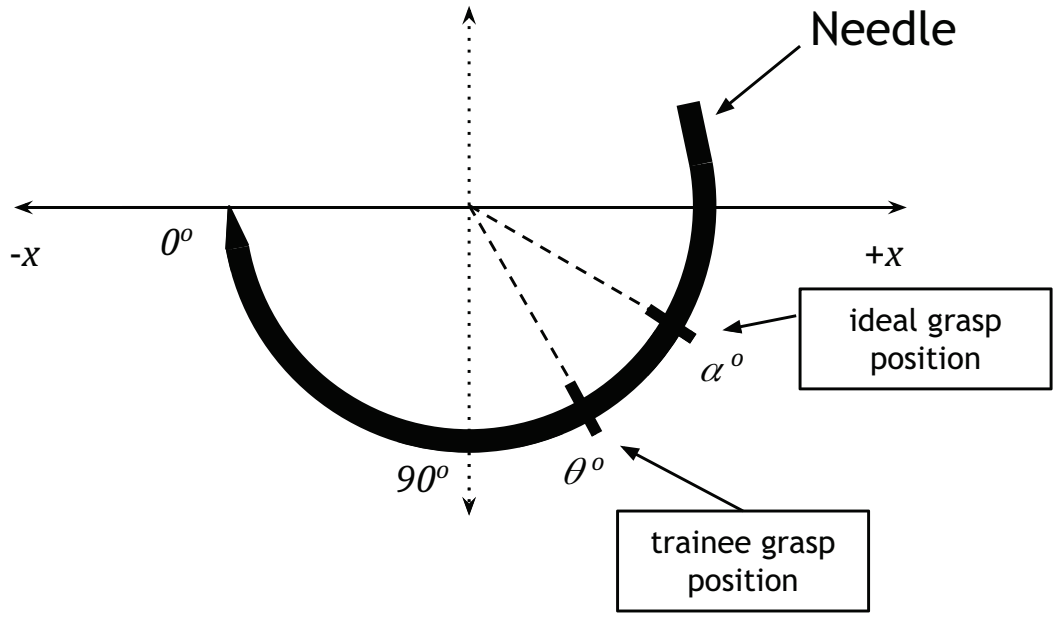


Figure 5.1: Grasp position metric: computed as deviation ($|\theta - \alpha|$) from ideal grasp position (150°)

close to the tip or the swagged end is not recommended as it can result in bending of needle, harm to tissue and awkward wrist orientation. SRSC modules suggest grasping between 1/2 and 2/3 of the length from the tip. However, this recommendation is for suturing a Penrose drain where the entry-exit targets are close spatially.

Assuming the midpoint of recommended interval is α, and the trainee's grasp position is θ, the metric value is calculated as $|\theta - \alpha|$ as shown in Figure 5.1.

Grasp Orientation: This is the mean of deviations from the ideal grasp orientation of the needle across all needle passing attempts. The recommended grasp is with the needle driver instrument jaws perpendicular to the needle plane (refer: SRSC Module 3 OSATS). Referring to Figure 5.2, a top view is presented with the needle along the

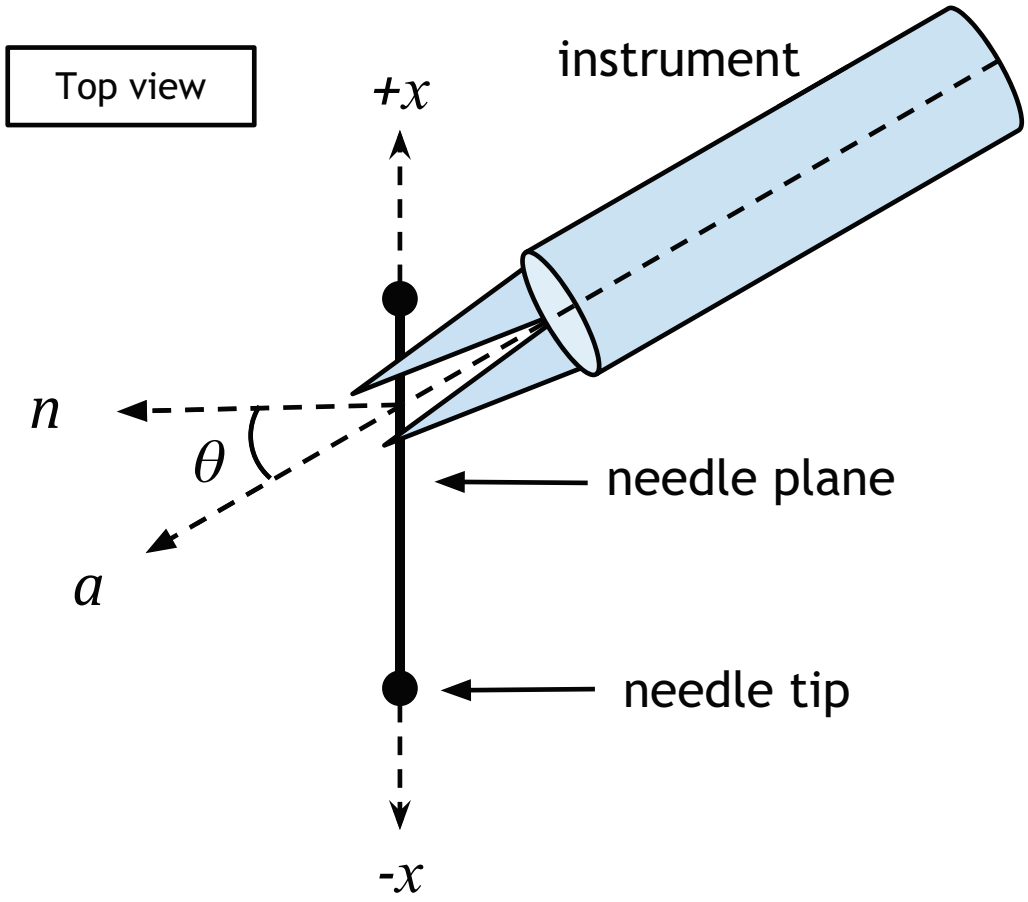


Figure 5.2: Grasp orientation metric: computed as deviation (θ) from ideal grasp along direction (n)

x-axis. n represents direction perpendicular to the needle, a is the direction of the trainee's grasp making an angle θ with n. In this case, the metric is simply θ.

Insert Position: This is the sum of deviations of needle insertion point from center of current entry target marked on tissue surface across all needle passing attempts.

Drive Orientation: This is the mean of deviations from ideal drive orientation across all needle passing attempts. It is recommended that while entering tissue

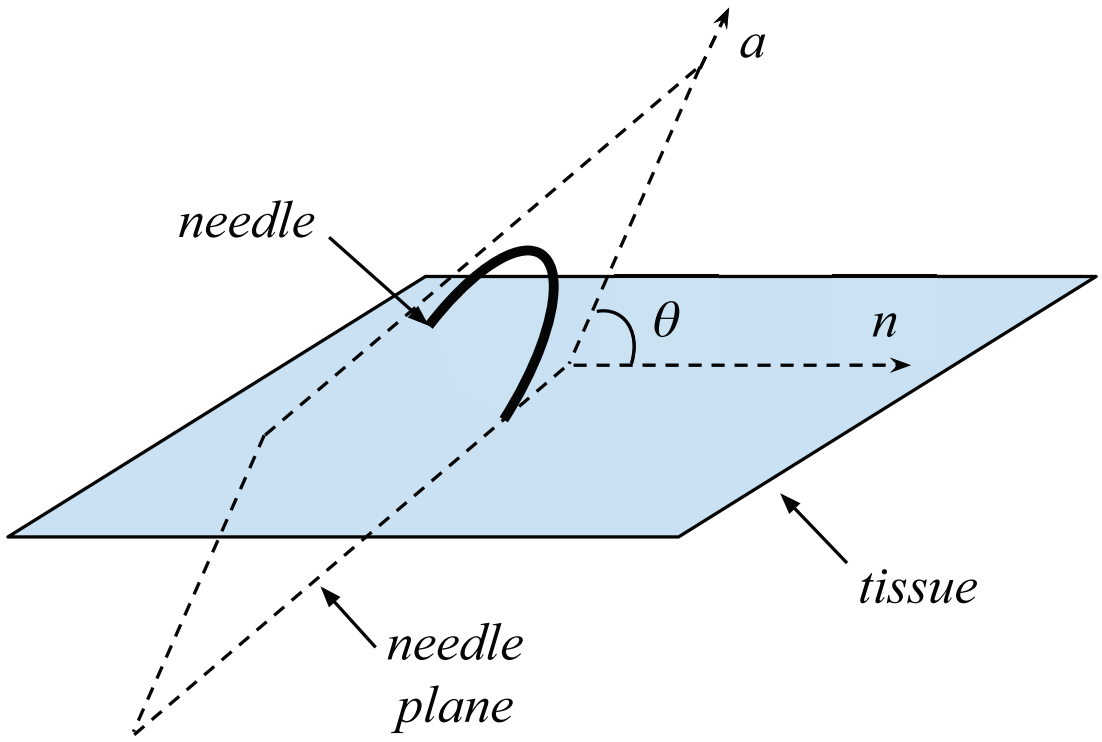


Figure 5.3: Drive orientation metric: deviation from perpendicular direction to the tissue surface (ideally θ =90°)

needle plane should be at right angles to surface (refer SRSC Module 3 OSATS). Figure 5.3 shows the needle plane oriented at an angle θ which can be computed using the directions a and n.

Drive Path: This is the mean of deviations of the needle tip from the ideal drive path through tissue across all needle passing attempts. Driving the needle along its curvature rotating one's wrist (or instrument wrist) is recommended practice SRSC Module 3 OSATS). The ideal drive path can be computed using positions of entry and exit targets and radius of curvature of needle as shown in Figure 5.4. This

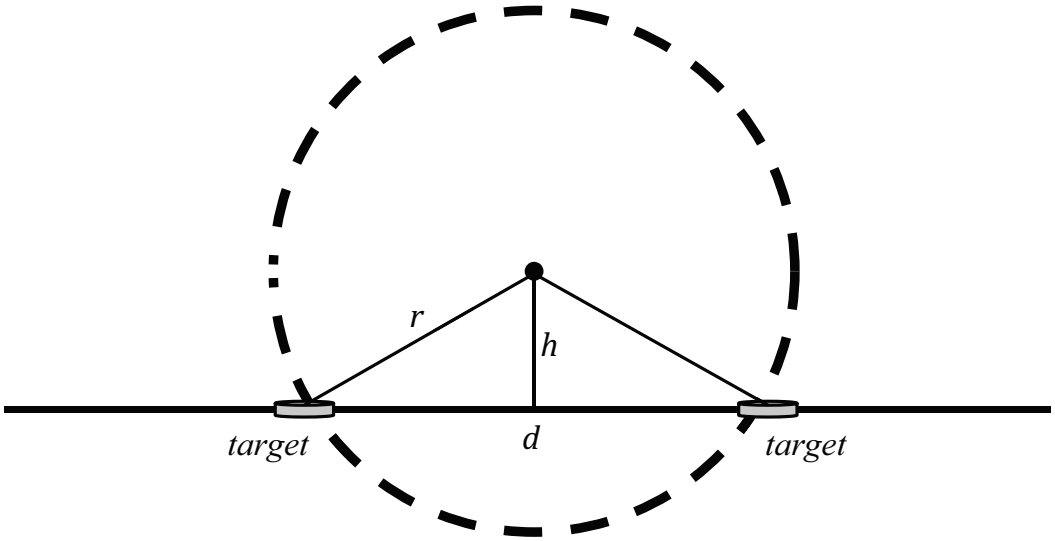


Figure 5.4: Ideal drive path (dashed circle) indicates the curvature of the needle with radius r. Given distance d between entry and exit targets, height h above tissue surface can be computed using Pythagoras theorem.

metric has two components – in-plane deviation (depth of the drive) and out-of-plane deviation (lateral shift from path) as marked by d_{in} and d_{out} in Figure 5.5. Each component itself is a sum of deviations along entire duration of the current needle passing attempt.

Exit Position: This is the sum of deviations of needle exit point from center of current exit target marked on tissue surface across all needle passing attempts.

Needle Grasps: This is the number of times the needle is grasped by either instruments across all needle passing attempts.

With the above metrics, we present Figure 5.6 that shows a task flow diagram for a single needle pass attempt along with trainee's actions and error and deficit metrics

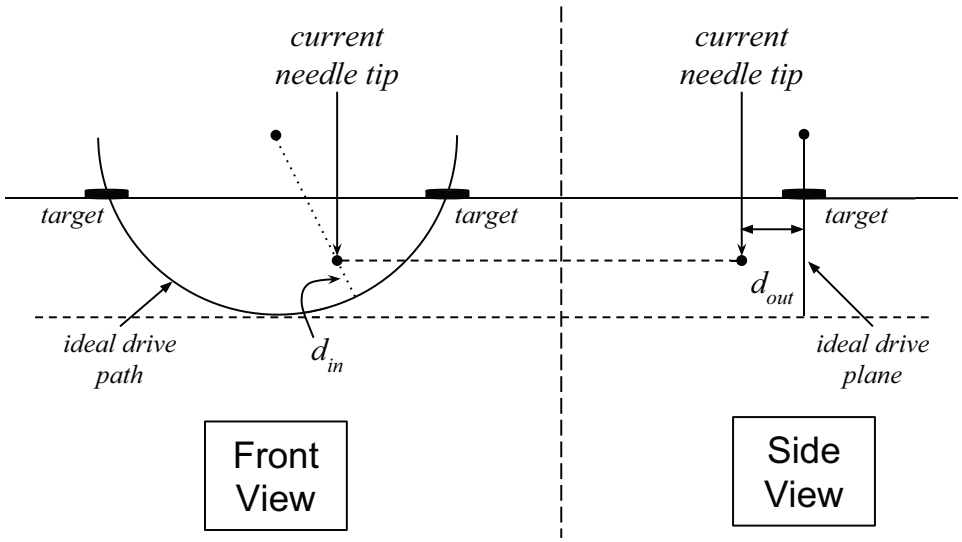


Figure 5.5: Drive path metric: deviation from ideal path defined by the needle curvature

associated to the action performed and state of the task.

5.2.3 Real-time Teaching Cues

We have presented error- and deficit-based performance metrics in the previous section. We highlighted that these measure the deviation from ideal performance and are associated with key learning elements of surgical skill. In case of needle passing (NP), loading (grasping) pose of needle with respect to instrument, insertion pose of needle with respect to tissue surface and path (trajectory) of needle tip through the tissue and out of it form the learning elements. While learning elements are important for skill evaluation, they form the basis for useful and individualized feedback delivery as well. We believe that a coach should provide criticism (feedback)

Figure 5.6: Error and deficit metrics for needle passing. Circles represent current state of NP. Solid boxes represent events like piercing, retraction and pull out. Thick arrows represent action performed by user to make progress and culminate in the events. Thin arrows represent logic flow based on success or failure of the progress attempt and result in state transitions. Error metrics are shown in dashed boxes. Deficit metrics are shown in dotted boxes.

and teaching (demonstration) at such learning elements to be effective. **With this belief, we propose the concept of real-time teaching cues as feedback and demonstration tools to deliver effective surgical skills learning.**

We will list and describe visual teaching cues for learning elements in the context of robot-assisted NP using the da Vinci® Skills Simulator™ (dVSim; Section 2.1.2). Figure 5.7 shows all of these cues during task performance. We will show zoomed in images of each cue in the respective sections below.

I. Ideal Instrument Indicator

What: This cue indicates the ideal instrument (left v/s right) for performing the current needle pass.

Cue: Red colored spheres at instrument tool-tip indicate the ideal instrument (Figure 5.8).

How: We use the instrument's initial setup pose, the current entry-exit targets pose and the joint limits on the robotic arms (PSMs) holding the instrument to determine the ideal instrument. In case of ambidextrous drives (left or right instrument may be used), we may base it of the handedness of the user which is obtained as an input while setting up the user's profile on dVSim.

Why: If a trainee selects the non-ideal instrument it can lead to constrained ergonomics, awkward needle insertion angles and can lead to unnecessary stress on tissue.

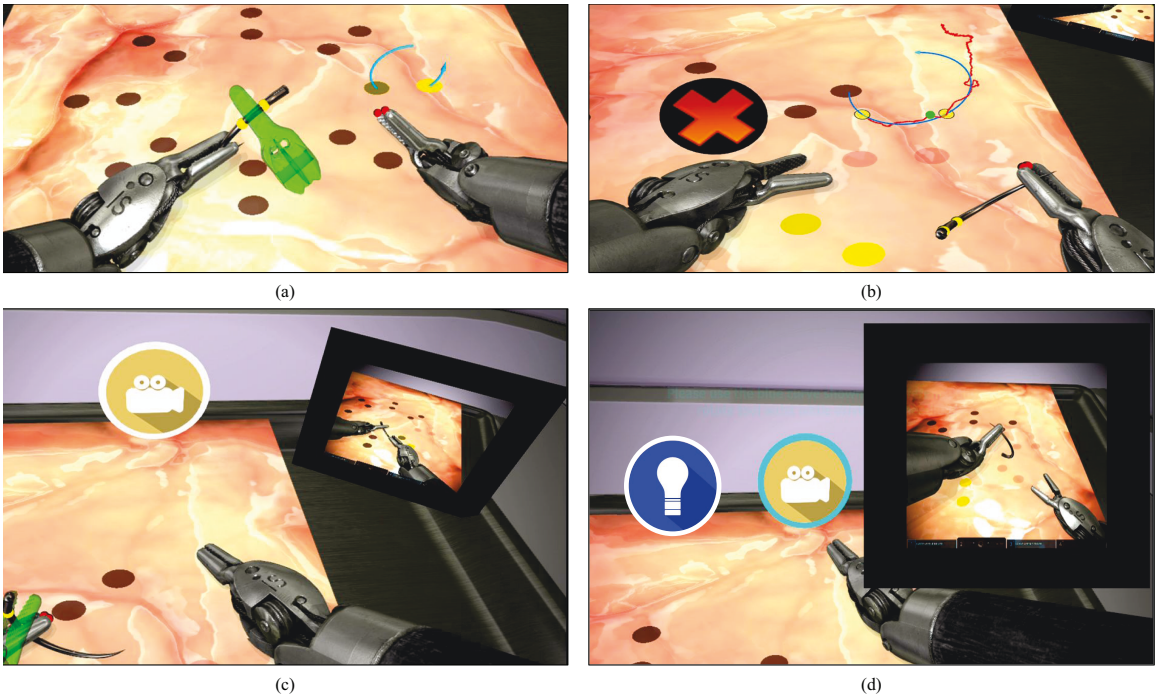


Figure 5.7: Real-time teaching cues using visual overlays in a robot-assisted VR needle passing task. (a) Red spheres on the right instrument tip are the Ideal Instrument Indicator, yellow spherical dots on the needle curvature are the Grasp Position Guide, semi-transparent green overlay is the Grasp Orientation Guide, cyan circular arc across the active targets (yellow) is the Ideal Drive Path Overlay; (b) Red curve trace is the Trajectory Playback Overlay and the green sphere indicates the needle tip position during the playback, red cross icon is the 'dismiss' icon to hide the trajectory overlay; (c) Video Demonstration Overlay on the side of the task environment along with the 'video' icon to toggle the video overlay onto the focal plane as in (d); (d) 'help' icon to enable/disable teaching cues in the 'User' mode.

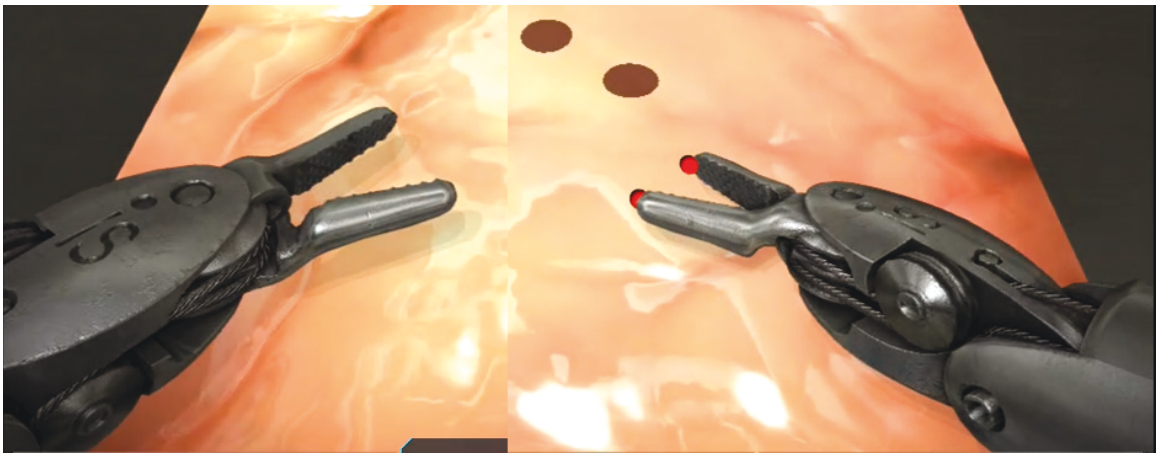


Figure 5.8: Teaching cue: ideal instrument indicator

II. Grasp Position Guide

What: This cue indicates the suitable range of grasping regions on the needle for ideal needle insertion and driving.

Cue: Flashing yellow spherical overlays along the needle curvature indicate the ideal grasp region (Figure 5.9).

How: If the tip of the needle is at the 0° location around the circle, the yellow spheres appear at 135° and 165° respectively along the needle curvature.

Why: Grasping the needle farther along its body allows the user to drive the needle through the insertion and exit targets in one smooth motion (*bite*). If the needle is grasped closer to the tip, it leads to unnecessary motion and force exertion on the tissue while trying to pass the needle through to the exit target.

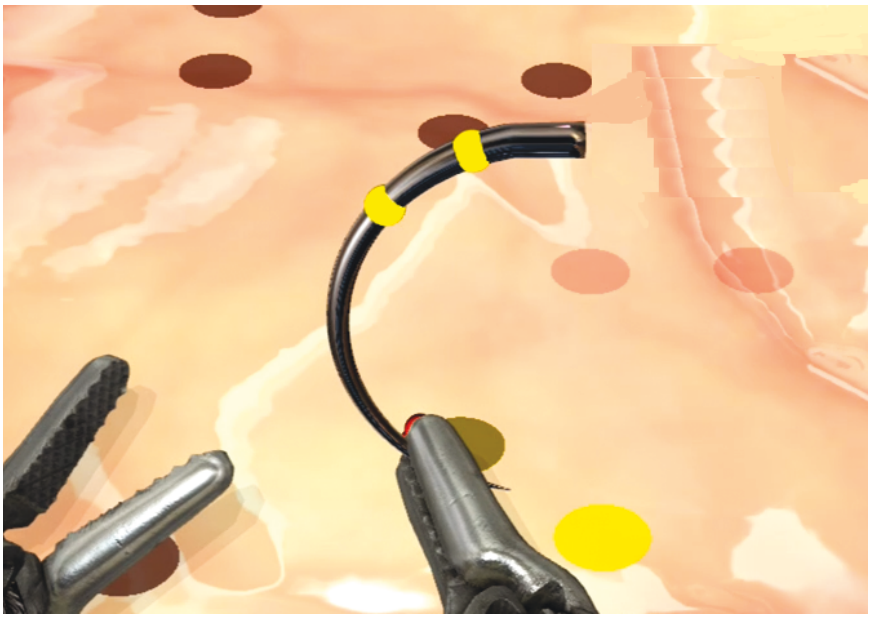


Figure 5.9: Teaching cue: grasp position guide

III. Grasp Orientation Guide

What: This cue indicates the ideal grasp angle of the needle with respect to the instrument.

Cue: A copy of the instrument's gripper graphics, however, with a light green semi-transparent appearance indicates the ideal grasp orientation (Figure 5.10). The transparency of the overlay increases as the ideal instrument is brought closer to the ideal grasp orientation, and it eventually disappears.

How: The overlay is positioned at the center (150°) of the ideal grasp position range. The ideal grasp orientation is perpendicular to the needle plane i.e. either along or opposite the needle plane's normal. Depending on the current needle drive direction (entry-exit target locations) and the ideal instrument suggested in Section 5.2.3, the

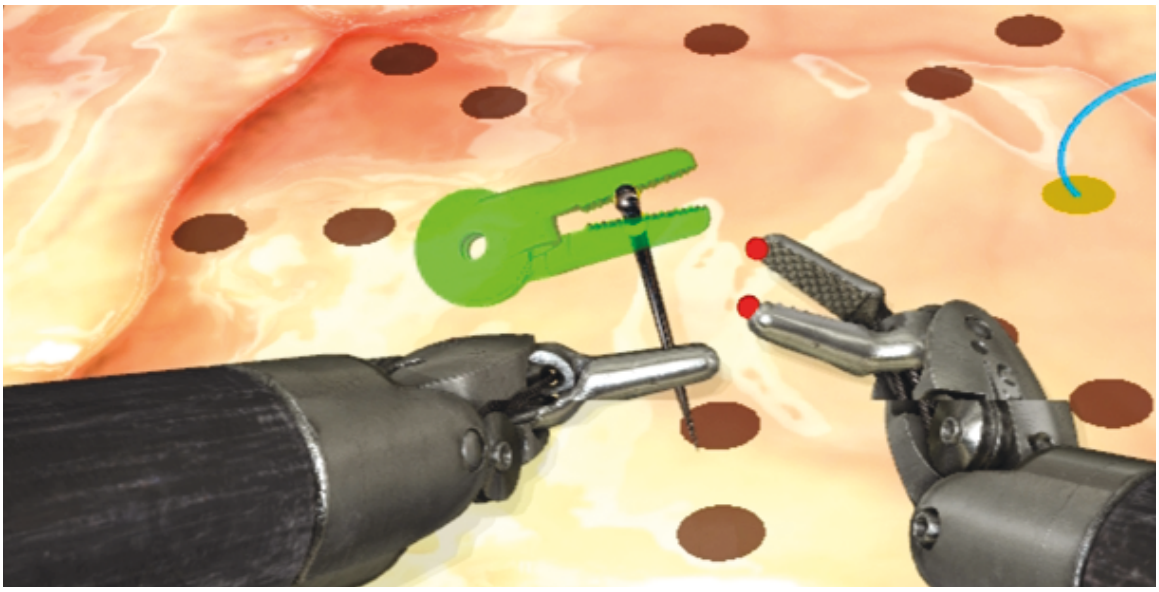


Figure 5.10: Teaching cue: grasp orientation guide

overlay is oriented along or opposite to the normal direction.

Why: The ACS/APDS skills module recommends grasping needle along the perpendicular direction for NP. Holding the needle in other orientations can result in excessive lateral force on the tissue at the insertion point, since the articulation of the instrument wrist required to rotate the needle through the tissue gets constrained.

IV. Ideal Drive Path Overlay

What: This cues indicates the ideal path for driving the needle through the tissue from entry to exit target.

Cue: Cyan colored arc passing through the entry and exit targets indicates the ideal drive path (Figure 5.11). An arrow pointing along the current drive direction appears at the end of the arc.

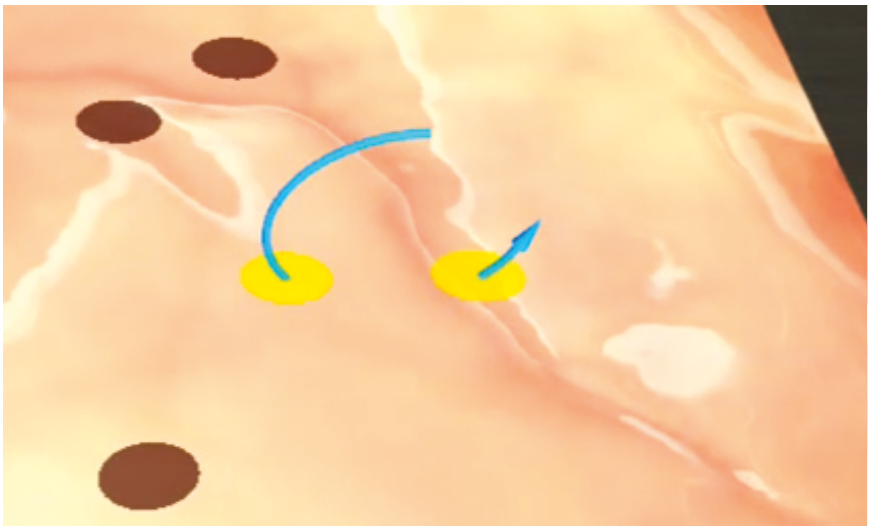


Figure 5.11: Teaching cue: ideal drive path overlay

How: Figure 5.4 shows the ideal drive path as the dashed circle. The ideal path has the same radius of curvature as the needle. The height of the center of the path (h) is calculated using the distance (d) between the entry and exit targets and needle's curvature radius (r). The orientation of the overlay is determined by the drive direction.

Why: This cue teaches the user to rotate the needle along its curvature, while driving it through the tissue as well as pulling it out. This rotational motion results in minimal lateral forces at both the insertion and exit targets as well as within the tissue compared to a straight motion drive. This is a recommended practice by the ACS/APDS skills modules.

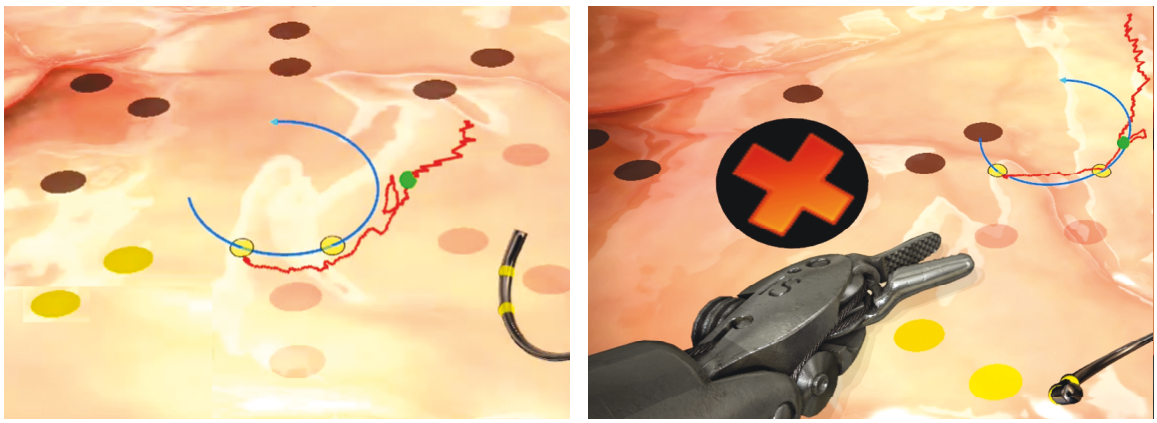
V. Trajectory Playback Overlay

What: This cue displays the trainee's needle tip trajectory from the previous NP attempt along with a graphical playback.

Cue: Red curve shows the needle tip trajectory along with the ideal path overlay (cyan arc) described in Section 5.2.3 (Figure 5.12a). The overlay appears at a certain height above the tissue surface and oriented facing towards the virtual endoscope for better visualization. A green spherical overlay runs along the trajectory to simulate the trainee motion as the drive was performed. Additionally, two flat yellow circles indicating the entry and exit targets are shown for reference. The overlay appears for a fixed period (10 seconds) during which the playback loops over. It is accompanied with a 'dismiss' icon that appears as a big red cross near the second instrument (the one that was not used to insert the needle; Figure 5.12b). The user can dismiss the trajectory overlay and proceed to the next NP segment before the 10 second period runs down by bringing their other instrument near this icon.

How: The trajectory of the previous needle pass is logged along with timestamps. The sphere's location is updated from the log. The ideal path overlay is computed as described in Section 5.2.3. The orientation of the cue is determined using the current endoscope view direction.

Why: Unlike the other cues, this cue is for reviewing one's performance. Such a playback cue that appears along with the ideal approach (drive path) gives an immediate visual feedback about deviation from ideal path and quality of the NP



(a) trajectory overlay cue **(b)** 'dismiss' icon to hide trajectory overlay

Figure 5.12: Teaching cue: trajectory playback overlay

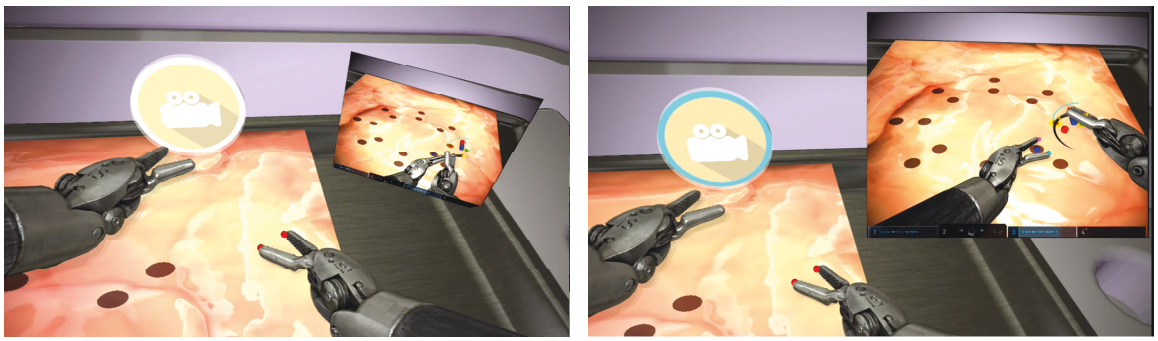
attempt.

VI. Video Demonstration Overlay

What: This cue shows a video playback of an expert/ideal performance of the current NP segment.

Cue: A movie texture appears in the simulation task environment to show the expert performance (Figure 5.13). The video is set to an infinite playback loop. Additionally, a 'video' icon is presented in the environment, so that the user can bring the movie texture onto the focal plane and in larger size to review the expert execution (Figure 5.13b). This icon is toggled by bringing either of the instrument tool-tips closer to the icon.

How: Pre-recorded expert (ideal) performances are stored for each configuration of NP segment. Movie texture's video is updated based on the NP segment that is being



(a) on the side **(b)** in view

Figure 5.13: Teaching cue: video demonstration overlay

attempted.

Why: This cue is important for showing beginners how the available instrument and needle should be manipulated to complete the NP attempt with success applying minimal lateral force on the tissue. The trainee can refer to expert videos while performing the task enabling real-time learning.

5.2.4 Real-time Coaching Modes

While teaching and feedback are important for learning skills, we believe that they should be adaptable to the trainee's current expertise level and needs. We propose three modes for real-time coaching that govern when and which teaching cues will be presented to the trainee for a more effective learning experience.

I. Teach Mode: The system shows all the teaching cues to the trainee throughout the task performance.

II. Metrics Mode: The system monitors current NP performance using metrics like

instrument path length, completion time, number of tissue pierces and force exerted on tissue by needle. Relevant teaching cues are displayed if the performance falls below par. A text message explaining which metric caused the cues to be triggered is displayed as well. The cues appear for that segment only.

III. User Mode: The trainee is provided with a 'help' (bulb) icon as shown in Figure 5.7d. The coach displays the teaching cues for the particular task segment if the trainee activates them by bringing either of the tools close to the help icon.

5.2.5 Teaching Cues and Task Progress

All the cues described so far appear at certain time points in the task flow. It does not make sense to show the trainee how to grasp the needle while they are already driving it through the tissue. Instead, it may result in distraction. Figure 5.14 shows the ON/OFF (visibility) state of these cues on the NP task timeline. Note, the figure is a mere illustration and duration for each action (arrows) varies based on trainee's performance. Cues of ideal instrument, grasp position and grasp orientation appear before the insertion of needle through the tissue. These cues guide the trainee to set up their instruments and needle in an ideal configuration for the current NP segment. Once the needle pierces the tissue, the ideal tool, grasp position and grasp orientation cues disappear. The ideal drive path is shown through out the setup phase and continues to remains in view during the driving phase to guide the trainee. Upon pulling out of the needle through the exit target, the ideal drive path

disappears and the trajectory playback overlay is shown for quick review of the NP attempt. A short delay is introduced in the simulation logic to let the trainee review. Post the delay timer, the setup cues appear again, but configured for the next NP attempt. The video demonstration overlay is always visible in the background of the simulation environment with an expert performance playing on a loop.

This logic is true for the Teach mode of coaching. Other factors determine the visibility in case of the Metrics and User modes. We shall postpone the discussion about that until later when we introduce the concept of *skill and coaching progression* in Chapter 6.

5.3 Randomized Controlled Trial

We conducted a randomized controlled trial (RCT) as a pilot user study to validate the effectiveness of real-time teaching cues on skill development in the needle passing task. Please refer to the description of ISI-SG-Sim Needle Passing data set in Appendix A.2 for study design, recruitment, data collected and other details regarding this trial.

We computed the change from baseline in performance metrics based on system events, instrument, console and endoscope motion, errors and deficits. We compared the control and experimental groups for performance improvement at the final task repetition using the Mann-Whitney U test. We also compared the improvement

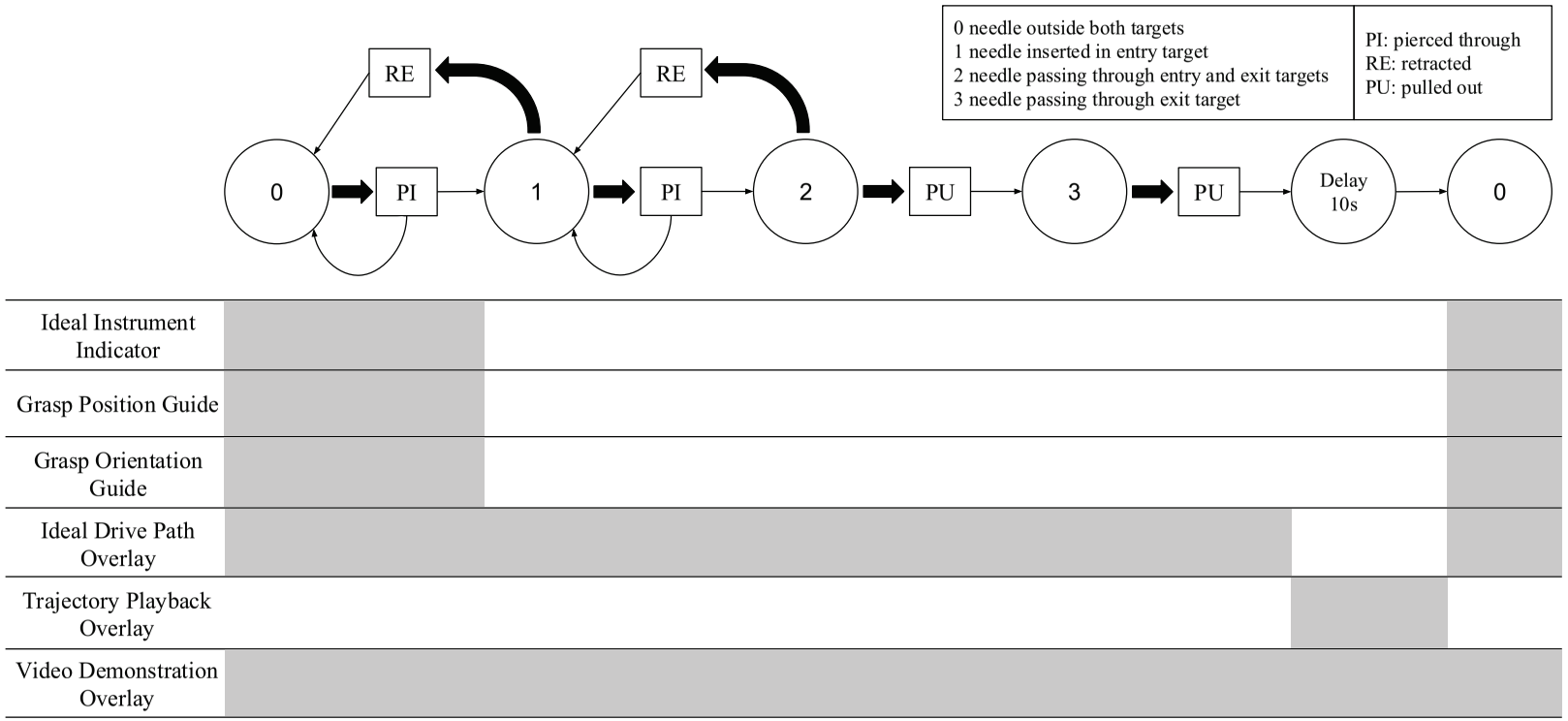


Figure 5.14: Needle passing timeline with visibility states of teaching cues. Gray regions indicate that the particular cue is visible to the trainee.

from baseline at the second, third and fourth task repetitions, since the experimental group was exposed to a new coaching mode each time. We tested the within group performance improvement for the control and experimental group using a Wilcoxon signed-rank test. We summarized responses from the pre- and post-study questionnaires as well.

Results

We assigned 16 participants to each arm; two participants in the experimental arm did not complete the study. Six participants had incomplete data due to technical reasons (two in experimental and four in control arms). We performed standard imputation for these incomplete data using mean for continuous measures and median for count-based ones. Finally, we analyzed data from 30 participants (control: 16, experimental: 14).

Table 5.2 reports demographics and prior experience of participants on da Vinci systems. We recruited 6 trainers (control: 3, experimental: 3) who are employed at Intuitive Surgical Inc. RAMIS training center (Sunnyvale, CA) where they orient and train surgeons on the dVSS. Remaining participants were engineers at the company with varying experience using the da Vinci systems and performing needle passing using them.

Table 5.2: Participant demographics and prior experience with da Vinci systems

Demographic	Experimental ($N = 14$)	Control ($N = 16$)
Experienced Trainer	3	3
Engineers	11	13
Right-handed operators	14	15
Experience in past month with:		
dVSS		
Never	1	3
1–10 hours	7	10
10–20 hours	2	1
>20 hours	4	2
dVSim		
Never	6	10
1–10 hours	5	3
10–20 hours	2	0
>20 hours	1	3
Other VR simulators		
Never	13	13
1-10 hours	0	1
10-20 hours	0	0
>20 hours	1	2
da Vinci-based needle passing		
Never	7	10
1-10 hours	4	3
10-20 hours	1	0
>20 hours	2	3

Post-study Questionnaire

In the qualitative survey, most participants felt the exercise was effective to teach needle passing on dVSim (80% answered 'effective' or 'extremely effective' and 20% chose 'neutral'). Most (93.3%) participants perceived an improvement in their performance relative to the baseline. Participants uniformly rated their baseline performance below par (poor: 33 %, below average: 33 % and average: 34%); they perceived their final performance as above average (excellent: 4%, good: 52%, average: 44 %).

In the experimental group, 78%, 57% and 50% of participants rated the *Teach*, *Metrics*, and *User* modes respectively as either "Useful" or "Extremely useful". 50%, 29%, and 21% of the participants in the experimental group preferred *Teach*, *Metrics*, and *User* modes, respectively.

Participants ($\geq$ 85%) rated all but one of the six teaching cues as intuitive, clear to understand, and effective for learning . 22% of participants in the experimental group found the trajectory playback teaching cue to be not useful, and the same proportion found it hard to understand and not intuitive. Some participants rated the text prompts appearing along with the teaching cues as unclear or not useful, whereas they found ($\geq$ 80%) the icons associated with the coaching modes and teaching cues to be clear to understand and easy to use. Finally, 92% of participants exposed to our VC framework in the experimental group felt that such feedback is essential for effective learning both in the presence and absence of a surgical educator or mentor.

In the control group, majority of the participants (68%) felt that real-time feed-

back would have helped them in improving their performance. While participants were equivocal about the effectiveness of real-time feedback, 93% of them preferred such feedback for the study task. In this sample of participants who were not exposed to our VC framework, 56% preferred the *Metrics* mode, 13% preferred the *User* mode, and 31% preferred a transition from the *Teach* to *Metrics* to *User* modes as they made progress in the skill level.

Task-level Performance

We observed statistically significant difference in the improvement of performance between experimental and control groups on one metric (Grasp Orientation Deviation) at the task level (Table 5.3).

Figure 5.15 shows the difference (effect size values) in task-level performance improvement over the baseline between experimental and control groups. Effect sizes[1] were calculated using Cohen's *d*. Time and motion efficiency metrics uniformly show a higher learning in control group (warm colors), while deficit and error metrics show higher learning in experimental group (cool colors). We observe that number of movements per second, deviation in grasp orientation and in-plane deviation from ideal drive path show statistically significantly higher performance improvement in the experimental group at repetition 2, 3 and 4 (movements) and repetition 2, 4 and 5 (grasp orientation deviation) and repetition 2 (ideal drive path deviation - in plane)

[1] `https://en.wikipedia.org/wiki/Effect_size`

Table 5.3: Performance improvement from baseline on overall task execution. Experimental and control group were compared using Mann-Whitney U test.

Metric	Experimental ($N = 14$)		Control ($N = 16$)		P-value
	Mean	Std Dev	Mean	Std Dev	
Completion Time	-132.71	134.05	-167.95	172.90	0.52
Path Length	-137.04	162.11	-208.81	227.47	0.13
Movements	0.25	0.39	0.03	0.45	0.14
Ribbon Area	-277.47	322.11	-427.91	453.19	0.15
Master Path Length	-337.42	376.86	-546.20	428.33	0.16
Master Workspace Volume	-294.24	528.96	-882.42	1403.51	0.04
Excessive Needle Pierces	-1.50	8.92	-5.88	14.57	0.14
Excessive Instrument Force (Count)	-1.79	2.86	-1.63	5.80	0.53
Excessive Instrument Force (Time)	-6.35	8.07	-5.72	19.79	0.47
Excessive Needle Tissue Force (Count)	-2.93	5.51	-4.44	15.59	0.97
Excessive Needle Tissue Force (Time)	-11.81	21.04	-17.70	48.45	0.76
Grasp Position Deviation	-3.73	16.86	4.19	18.32	0.31
Grasp Orientation Deviation	**-14.53**	**12.99**	**-4.22**	**11.09**	**0.04**
Ideal Drive Path Deviation (In Plane)	-0.00	0.06	-0.01	0.05	1.00
Ideal Drive Path Deviation (Out of Plane)	-0.01	0.05	-0.02	0.03	1.00

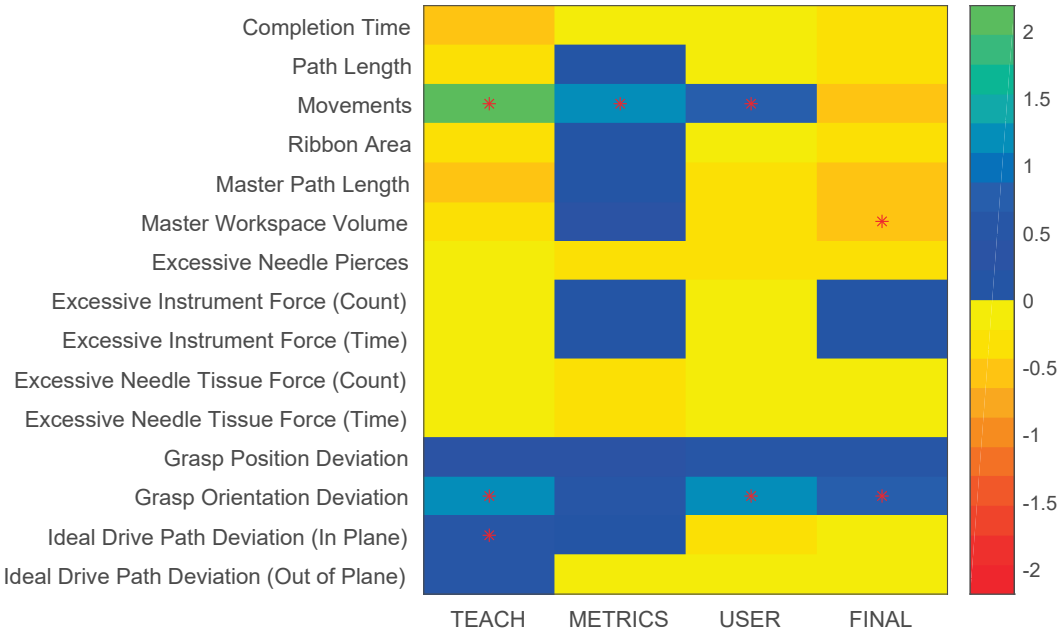


Figure 5.15: Comparison of experimental and control groups at each task repetition for task-level performance improvement over baseline. Each cell is the effect size value for the metric. Negative values (warm colors) indicate larger improvement in control group and positive values (cool colors) indicate larger improvement in experimental group. Red asterisks represent that the P-value was less than 0.05 for particular metric and task repetition.

respectively. Also, experimental group shows higher improvement in the number of movements in the TEACH mode repetition which becomes smaller than the control group improvement by the FINAL repetition. Deficit metrics (lower four in the figure) indicate higher learning in experimental group in the TEACH mode. This learning reduces compared to the control group by the FINAL repetition.

We show within group task level performance improvement (effect sizes are used for scaling different metrics to a unitless quantity) in Figure 5.16 (experimental) and Figure 5.17 (control). Effect sizes were computed as the fraction of sample mean and

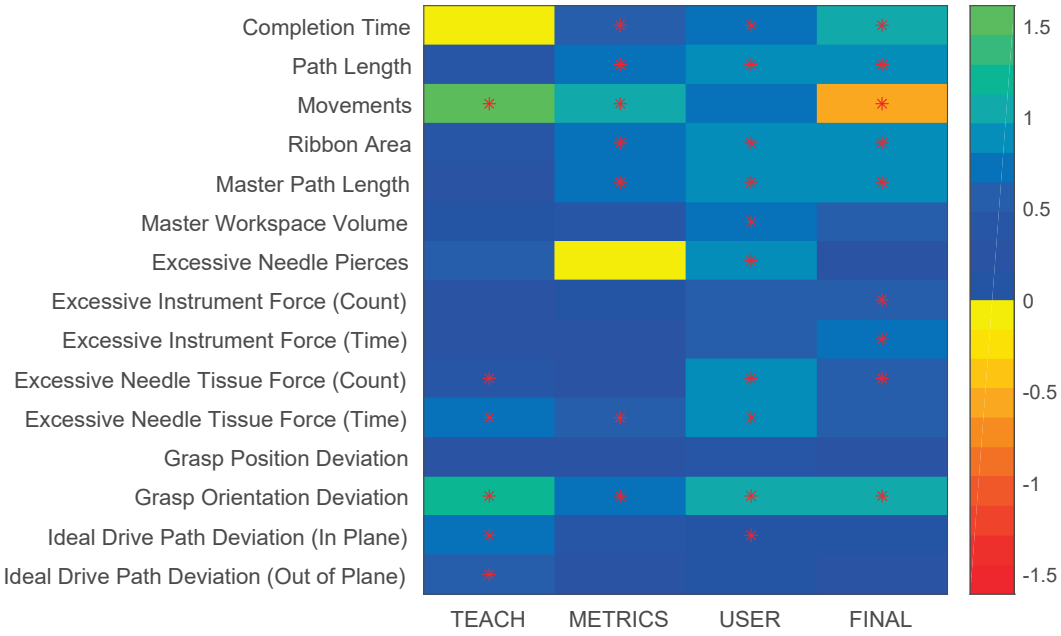


Figure 5.16: Task-level performance improvement over baseline within the experimental group for each task repetition. Each cell is the effect size value for the metric. Negative values (warm colors) represent skill deterioration and positive values (cool colors) represent skill development. Red asterisk represents a P-value less than 0.05 for the corresponding metric and task repetition. Task repetitions were in different coaching modes: 2 (Teach), 3 (Metrics) and 4 (User)

sample standard deviation. We observe that the experimental group shows significant performance improvement in error and deficit metrics initially (TEACH mode), and time and motion efficiency metrics later on (repetitions 3 to 5). The performance improvement in deficit metrics becomes insignificant at later repetitions except for deviation from grasp orientation. In case of the control group, we see almost no significant improvement in deficit metrics across all repetitions. However, the control group shows significant improvement in time and motion efficiency metrics throughout the task repetitions as well as the excess needle force (time) error metric.

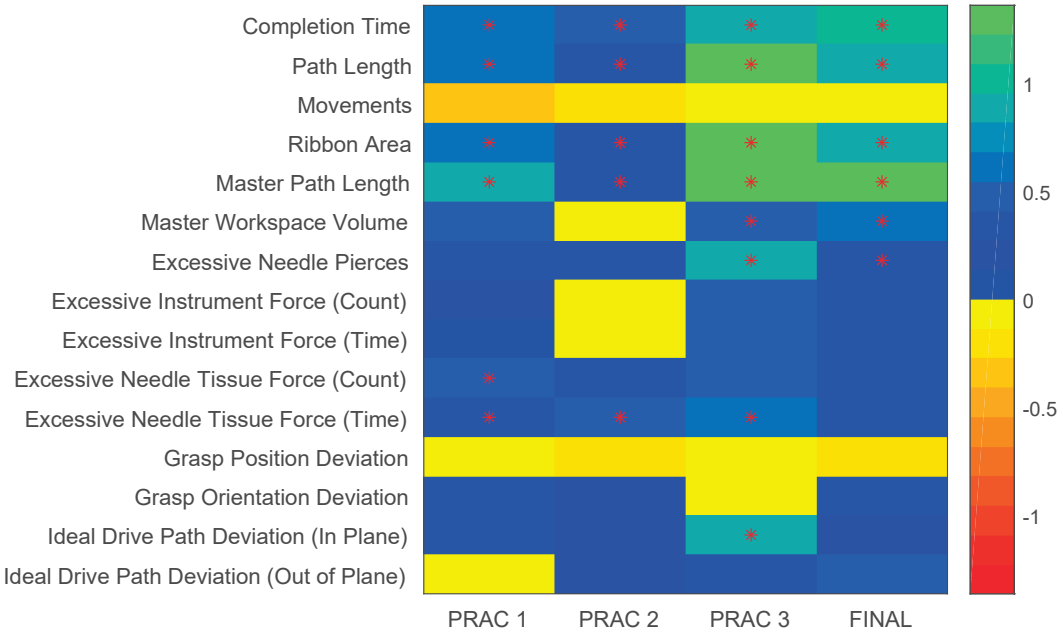


Figure 5.17: Task-level performance improvement over baseline within the control group for each task repetition. Each cell is the effect size value for the metric. Negative values (warm colors) represent skill deterioration and positive values (cool colors) represent skill development. Red asterisk represents a P-value less than 0.05 for the corresponding metric and task repetition.

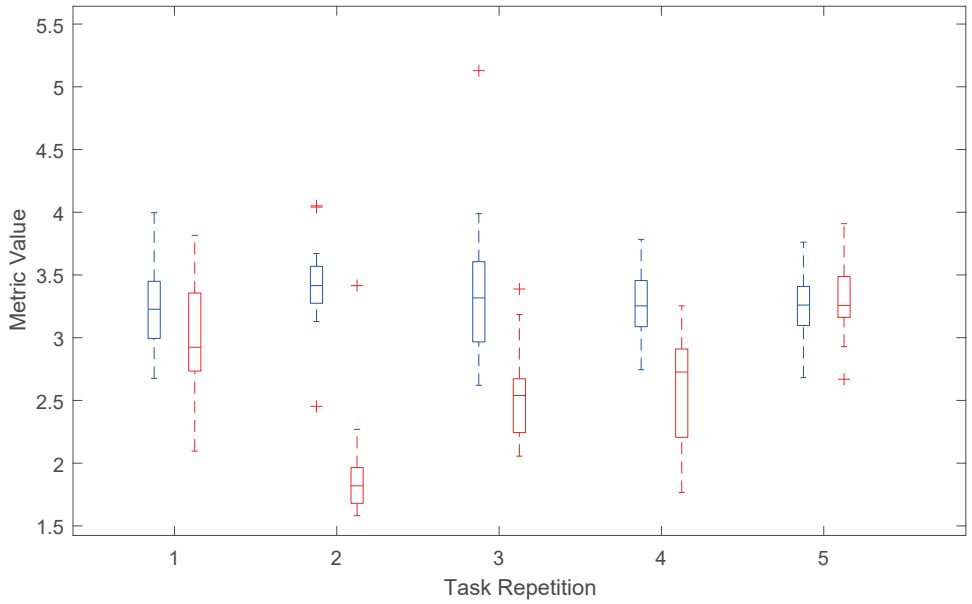


Figure 5.18: Number of movements per second for control (blue) and experimental (red) groups

Figures 5.18, 5.19, and 5.20 show the task-level performance for number of movements, deviation in grasp orientation and in-plane deviation from ideal drive path. We can see that the two groups were similar in their baseline performances. We confirmed this by performing a Mann-Whitney U test and saw that none of the metrics showed statistical significance ($p \geq 0.5$) except that of excessive exit needle pierces. These plots narrate similar story of higher learning in experimental group at TEACH mode and a return to the same level of performance for both groups at FINAL repetition.

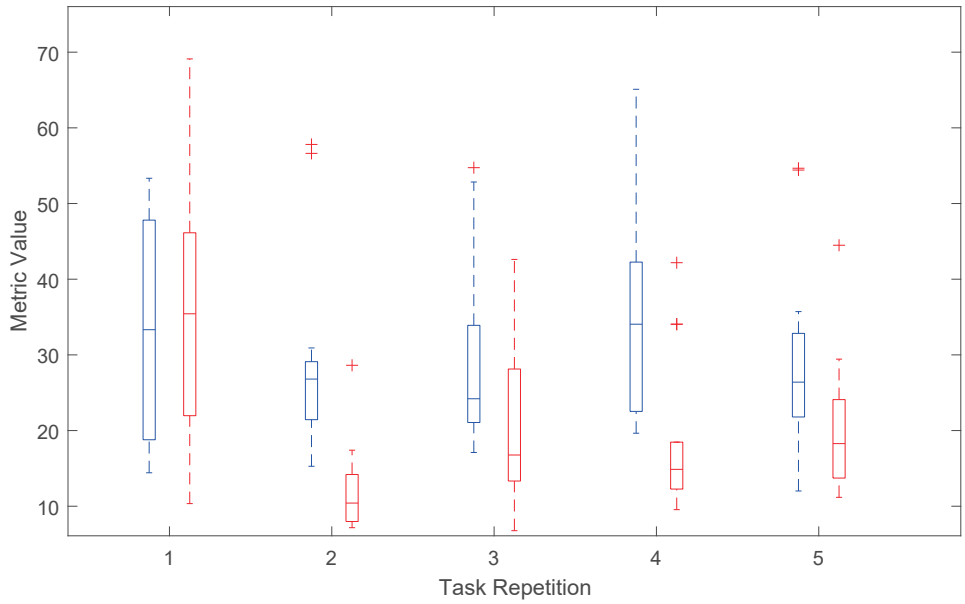


Figure 5.19: Deviation in grasp orientation for control (blue) and experimental (red) groups

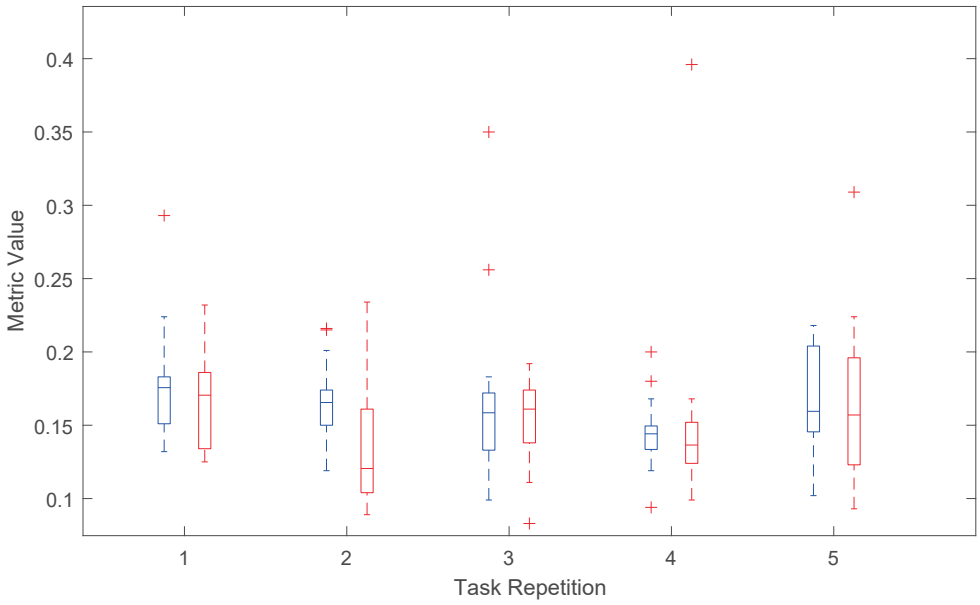


Figure 5.20: In-plane deviation from ideal drive path for control (blue) and experimental (red) groups

5.4 Discussion

Our finding of no statistically significant improvement in motion efficiency between the experimental and control groups is consistent with observations in previous RCTs studying feedback and coaching.[46,70,78,197,198] The real-time teaching cues were targeted at learning elements of the task. And, we did observe (Table 5.3) statistically significant improvement in deficit metrics (grasp orientation deviation) for the experimental group compared to the control group. The deficit metrics are crucial and relate to the product quality of the task, while motion efficiency metrics are meaningful once the task is completed with competent outcomes. However, the length of our study limited the learning experience for the participants. They might not have reached a plateau on their learning curves after just three task repetitions. We think that a future study with more number of task repetitions will show significantly better performance in motion efficiency, errors and deficit metrics for trainees receiving feedback and teaching using our framework.

With regards to coaching modes in our study design, there were two limitations. First, we exposed the experimental group to a single TEACH mode training session. We see significantly higher learning in the experimental group for the TEACH mode, but this learning becomes less significant for the later task repetitions (Figure 5.15). This learning behavior within the experimental group can be seen in Figures 5.18, 5.19 and 5.20, where the metric scores improve (decrease) significantly for the second repetition and then become worse (increase) for repetitions 3, 4 and 5. This may be

due to a short teaching intervention (lasting only one repetition). It indicates that the framework was successful in imparting knowledge about ideal needle passing skills to the participants in the experimental group, but a single teaching intervention was not enough to retain the knowledge. Second, we exposed the experimental group to all three coaching modes one after the other. These modes correspond to skill progression and autonomy of the participant in executing the task, and were not suited for the current short length study design. Future study design should inject these modes as separate arms with a larger participant pool and longer length of intervention to understand their effectiveness separately in skill development.

Additionally, we compared the performance improvement for the sub-group of participants who were *trainers* at Intuitive Surgical Inc. We did not observe any significance in their performance improvement across all the metrics irrespective of their randomization. This was expected since the current cues are targeted at teaching the skill of needle passing to beginners. Perhaps, a future extension of the framework to provide deliberate practice might be more useful in case of experienced users training with the goal of furthering their existing proficiency.[46,199]

The perspectives of the participants as indicated by the post-study questionnaires showed value for such teaching cues in being effective tools for skill development. All of the cues were rated highly in terms of clarity and usefulness. The lower favoring for trajectory playback overlay could be due to lack of clear instructions explaining the cue. In cases where the trainee struggles a lot to get the needle through the

exit target, the trajectory overlay can be overly crowded and prove to be negative in imparting knowledge or review. A better review cue or demonstration cue might help resolve this effect. The text prompts also received negative ratings. One reason for this would be that they appear a bit away (on top of the stereo viewer) from the focus area (middle of the screen) of the trainee. Trainees may not see them in time or just ignore them compared to the visual cues that were placed right at the point of action.

We explored visual teaching cues in the current work. However, other modes of feedback have been studied in prior work related to surgical interventions like audio, haptics and force sensing. In future, haptics-based cues that use virtual fixtures[15,192] can be added to provide a hand-over-hand guidance to teach expert behavior at learning elements in task as well as to give feedback at segments of sub par performance. Similarly, an immediate scoring framework with segment-level assessment of skill that informs trainee about current proficiency might be an effective cue as well.

Our objectives for the pilot study were to demonstrate the real-time feedback and its feasibility. In addition to achieving these objectives, this has provided estimates of variance needed to calculate the appropriate sample size for a subsequent definitive study on the effectiveness and value of automated coaching in surgical training. This future study should target the current limitations – non-surgical participants, shorter duration of practice, and single study arm for different coaching modes. Finally, the concept of teaching cues and deficit metrics should be extended and validated for

other surgical skills tasks as well.

5.5 Summary

In this chapter, we surveyed previous studies that have shown the effectiveness of delivering expert feedback and demonstrating ideal performances on skill acquisition and development. Similarly, we presented works from the literature that have shown value of teaching and identifying errors for improved learning. With this motivation, we proposed and implemented a framework to deliver such error-based feedback and teaching in a virtual reality setting.

Previous methods have relied on the availability of an expert to deliver effective teaching and feedback. The scalability of such an approach is limited and has motivated the need for automated, objective and individualized feedback and demonstration. Virtual fixtures-based techniques to guide trainees onto correct instrument motion paths have shown success but do not provide guidance on errors and deficits. We have demonstrated the feasibility of an automated framework to deliver real-time teaching and feedback on a commercial training platform for the first time.

Specifically, we introduced learning elements to deliver feedback and teaching at critical and consequential elements of a surgical skill. We introduced teaching cues to demonstrate ideal behavior of learning elements. And, we introduced error and deficit metrics to provide feedback on deviations from such ideal behavior at learning

elements. We explored these concepts in the context of robot-assisted needle passing. We described the study design and outcomes of a pilot randomized controlled trial that was conducted to study the effectiveness of these teaching cues in learning the needle passing task on a VR simulator (dVSim). We laid out future steps towards a larger study and a better teaching framework.

We have demonstrated the feasibility of automated and real-time feedback and demonstration for the first time in a VR simulation training setup.

Chapter 6

Towards an Automated Virtual Coach

With the increased restrictions on resident duty hours, concerns regarding learning in OR on patients and pressure on surgeons to increase OR efficiency, surgical societies felt the need for development of structured training curriculum for development of residents' surgical skills outside the OR. Programs like Fundamentals of Laparoscopic Surgery (FLS)[4] were created for certification of surgeons to ensure proficiency. Although, the guidelines require re-certification only every 7 to 10 years. In any other profession where performance is critical, such a long period for proficiency re-assessment is unheard of and considered as good as unreliable proficiency. For example, first class air pilots' certificates expire every year and every six months if the pilot is older than 40 years (DOT Federal Aviation Administration). However,

even if surgery programs mandated a re-certification every year, the current resources are nowhere close to what would be needed to make that happen.

Surgical coaching is as an alternative to regular standardized testing for proficiency development. Coaching is much more beneficial than standardized testing due to its inherent individualized nature.

6.1 Background

Let us look at some of the works in this recently popular, and active research field of surgical coaching.

6.1.1 Deliberate Practice

Practice maketh a man perfect

Ericsson[200] suggested that expert performance is not the outcome of extensive experience alone, but a result of deliberate practice. Deliberate practice (DP) is defined as "activities that have been specially designed to improve the current level of performance" by Erisson et al.,[201] in contrast to work (constrained activities motivated by reward) and play (activities that are enjoyable, with no goal). They argued that "differences between expert performers and normal adults reflect a life-long period of deliberate effort to improve performance in a specific domain." In a review piece,[199]

Erisson presents empirical evidence "for and against the presumed congruence between level of socially ascribed expertise [in quotes] and performance in medicine".

Few studies have looked at the effectiveness of DP in the domain of surgical skills and expertise. Crochet et al.[46] showed that DP enhanced surgical performance quality and resulted in superior transfer of skill onto real tissues compared to a conventional training group. Similar results were found by Hashimoto et al.[75] suggesting that DP leads to higher quality performance in VR laparoscopic cholecystectomy compared to standard training.

DP relies on the ability to assess performance and provide feedback. While this might be possible in an independent learning mechanism, previous work[76] has shown that physicians have a limited ability to accurately self-assess. Also, previous literature showed external feedback-based learning to be more effective in skill learning compared to the independent learning setting (Section 5.1). This leaves no option but to use faculty surgeons' (limited available) time to deliver DP by reviewing trainee performance and providing feedback – live or via video recordings. Hashimoto et al.[75] bring out this as one of the big challenges facing the success and inclusion of DP in surgical training curricula.

We believe that automated DP can be implemented in VR simulation setting.

6.1.2 Manual Coaching

While previous works have described concepts similar to that of a "coach", the popularity of surgical coaching came around the time Gawande[8] wrote in *The New Yorker* about the idea of a coach in the OR. Since then, several works have proposed surgical coaching frameworks and validation trials in various settings[9] showing the value of coaching for achieving surgical proficiency and guiding continuous professional development effectively.

Some of these works were observational studies. Birch et al.[202] proposed a new comprehensive teaching concept for surgeons in practice using mentoring. They showed it to be an effective strategy for safely introducing MIS into practice with an increase in total volume of MIS cases and decrease in total conversions to open surgery as well as decreased intra-operative complications. Briët et al.[203] showed the feasibility of an on-site coaching and monitoring by an experienced visiting surgeon for gynecologists. Hu et al.[204] reported a qualitative study on video-based coaching. They recommended characteristics for coaching programs and observed that surgeons at all levels found such video-based postgame analysis to be useful and highly instructive.

A handful of control trials have been conducted to test the effectiveness of coaching:

1. Cole et al.[198] compared the effects of structured coaching and autodidactic training in VR laparoscopic cholecystectomy surgery. They observed that pro-

ficiency indicators including error reduction, understanding of surgical strategy and surgical quality improved significantly with the group receiving structured coaching compared to the group that received no additional training.

2. Bonrath et al.[47] also compared individualized coaching based on the PRACTICE model[205] with conventional residency training by measuring post-intervention technical performance improvement over baseline in bariatric (jejunojujunostomy) surgery in the OR. They showed improved technical skill scores and reduced number of errors as well as enhanced self-assessment skills in the coaching group.

3. Karam et al.[206] compared one-on-one video-based coaching review of resident's performance in simulated fluoroscopy guided articular fracture surgery to no coaching. They showed significant improvement in OSATS scores and reduction in fluoroscopy utilization for the coaching group.

4. Singh et al.[78] investigated whether video-based coaching using the GROW model[207] enhanced laparoscopic skills performance in VR and porcine cholecystectomy. They showed higher GRS was observed for coaching group in both cases.

5. Palter et al.[208] assessed the efficacy of peer-based coaching for teaching laparoscopic suturing in simulation to inexperienced faculty surgeons. They observed that the intervention group showed higher improvement in technical proficiency.

6. Soucisse et al.[209] compared video-based coaching using the GROW model[207] for resident's performance on side-to-side anastomosis in a cadaveric dog bowel compared to conventional resident training. They showed that coaching resulted in significantly higher increase in OSATS scores compared to no coaching. The residents reported that video coaching was a time-efficient teaching intervention.

In addition to these validation studies, Greenberg et al.[48,210] (Wisconsin Surgical Coaching Framework) and Stefanidis et al.[211] (Carolinas Coaching Model) have detailed characteristics and needs of a coaching framework along with recommendations for future validation of surgical coaching. They are both targeted at technical, cognitive and non-technical skills (Greenberg et al. make a distinction between cognitive and non-technical skills while Stefanidis et al. combine them under non-technical skills). Greenberg et al. divided coaching activities into three domains: setting goals, encouraging and motivating, and developing and guiding. While, Stefanidis et al. developed a 5-step coaching model: assess skills through video, identify areas of growth, group review and individual feedback, deliberated practice with coaching, and monitor patient outcomes.

While all the above works have proved the effectiveness of coaching for continuous professional development as well as beginner skill acquisition, they listed a common limitation with respect to surgical coaching – time. Both, coach's and coachee's time are valuable and constrained. Unlike other coaches in athletics or gymnastics, surgeons have a lot of other duties to perform that include operating patients, attending

clinic, writing grants and managing clinical administrative paper work. Taking time out to provide coaching to residents or peers becomes of lowest priority. Unlike lecturing, coaching requires individualized attention and face-time to evaluate, guide, critique and mentor the coachee. This restricts the flexibility with which coaching can be adopted by surgeons and residents in their existing busy schedules. All the listed studies were limited to a single or handful of coaches, which is another deterrent factor in large scale introduction of coaching in its current form. Rooney et al.[212] compared faculty ratings between live versus video-recorded resident performances and faculty versus skills coaches' ratings of video-recorded resident performances. They found that trained skills coaches may be used to reliably assess video-recorded performances to make better use of available personnel and minimize the time surgical faculty need to be present in the skills lab for assessment purposes. They suggested that use of video-recorded performance ratings and skills coaches may be viable alternatives to live ratings performed by surgical faculty.

Surgical coaching, however, still faces a lot of other challenges. As Mutabdzic et al.[7] mention, the surgical culture values the portrayal of competency and instills the value of surgical autonomy. They state that "coaching, in its traditional sense, cannot be achieved with the surgeon as learner [in quotes] and having full control". Additionally, they found that the surgeons considered proficiency in technical skills beyond a certain level is not as important as other requirements that may be considered in annual performance reviews. Finally, coaching can be subjective and biased

by the coach's viewpoint compared to the coachee's.

An automated coach can resolve all of these limitations while delivering efficient and effective feedback, detailed evaluation, relevant demonstrations and individualized deliberate practice sessions. Concerns related to portrayal of incompetency and time resources are dealt with, since automated coaching can be done in a private setting and at the flexibility of the coachee surgeon. Subjectivity, availability and financial burden are taken care of in an automated coach setting, as well.

As a result of the growing recommendation for surgical coaching and the listed obstacles in delivery of manual coaching, we have undertaken the work presented in this thesis so far, towards the development of an automated virtual coach (VC) for surgical skills acquisition, development and retention.

We will present the VC framework in the following sections. We will list and describe the coaching activities that the VC should deliver upon. Then, we will present a system diagram of the framework and explain its core components. Finally, we will present the software architecture used to implement the VC and end with future work towards additional features, validation and testing.

6.2 Coaching Activities

We propose the VC with a 5-step activity model (Figure 6.1) that is in line with coaching activities other works have observed or proposed.[45,210,211] For any skill that is do be developed or improved by a coachee, the coach must deliver interventions using all of these activities. We will now explain each activity in the context of a scholastic setting and draw analogues to surgical training. By doing so, we hope to explain the generalization of the VC to other learning environments outside of surgery. Consider the scenario of students (*coachee*) in a course of calculus learning differentiation (*skill*). Let us assume that the differentiation module is structured so that the students are assigned regular tests (*performance*).

6.2.1 Demonstrate: *how do I do it correctly?*

The teacher (*coach*) outlines the concept of differentiation (assuming the students have the pre-requisite knowledge) and provides examples of how to differentiate elementary mathematical functions (*demonstrate*). The teacher may also show some tips and tricks to perform differentiation with ease. Similarly, a surgical coach must *demonstrate* the ideal approach to perform the surgical skill. The coach may use *teaching cues* (Chapter 5) to show approaches for expert-like execution of the skill. Additionally, the ideal teacher provides an answer key to the test questions for the students to know what was the correct approach to solve the differentiation problem.

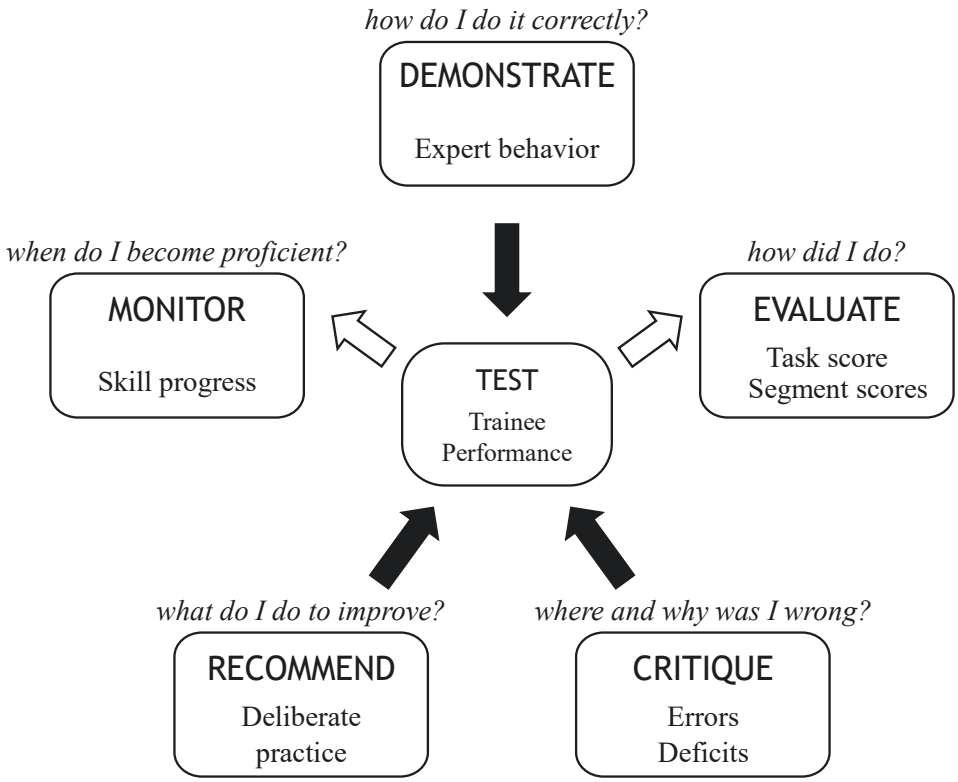


Figure 6.1: Coaching activities of our automated virtual coach (VC)

The VC must provide relevant expert performances to the coachee as well.

6.2.2 Evaluate: *how did I do it?*

After the students submit their test solutions, the teacher must grade (*evaluate*) them and give back the grades before the next test for the students to reflect upon their performance. Irrespective of what criterion are used to perform this evaluation, the teacher has to choose a standardized and objective method to do so. Likewise, the VC must evaluate the surgeon's performance and provide an objective score to them in time before their next performance else such evaluation may not be effective.

6.2.3 Critique: *where and why was I wrong?*

Imagine the scenario wherein the students are given back their graded solutions but with a single grade letter (*task score*) on top and no other information about it. This would hamper the student's learning. However, this is the current scenario of surgical skill evaluation. Luckily, our teacher is a good coach and provides a question-by-question breakdown of the grade (*segment score*), and the students can know where to look for mistakes. The VC must evaluate surgeon's performance at the task and segment levels (like the pairwise comparisons-based score presented in Chapter 4).

Even after knowing where the student lost their grade points, he/she may not

know what was wrong in their solution to the question. The ideal teacher redresses this issue by writing notes (*critique*) or annotating (circling) the mistakes (*errors and deficits*) for each grade point that was deducted. The VC must critique the surgeon's performance and provide information about errors and deficits that occurred during the performance (like the error and deficit metrics introduced in Chapter 5).

6.2.4 Recommend: *what do I do to improve?*

The teacher assigns homework (*deliberate practice*) to students that is directed at the concepts taught in class that day or week. Sometimes homework may be targeted at concepts on which majority of the class scored low grades. The teacher may provide extra sections (after school) for students to work on certain types of questions or concepts that may not be clear to them in an *individualized* setting. The VC must *recommend* focused practice sessions based on the surgeon's past performance evaluation and critiques. These practices must be deliberately targeted at low scoring segments and frequently occurring errors.

6.2.5 Monitor: *when do I become proficient?*

The teacher must maintain a record (*monitor*) of grades assigned to students on tests as well as homework completed by them to estimate their learning and graduate (*progress*) them to the next concept or course or grade. The teacher must also monitor

the class grades to adapt the teaching style for a better learning experience of the class. Slightly different, however, the VC must track the surgeon's progress on the skill and adjust its behavior and skill difficulty level to provide efficient learning. The VC must monitor the performance score history to recommend deliberate practice as well.

6.2.6 Online Coaching

Some of the activities described above (*demonstrate, evaluate and critique*) may be in play during the performance. A tutor working on a differentiation problem with the student may assess the correctness of the solution (*evaluate*), prompt the student about a mistake (*critique*) and provide hints to correct the solution (*demonstrate*) while the student is still working on the solution. Similarly, the VC can perform these activities in an *online* mode (similar to concurrent feedback or real-time teaching in Chapter 5) measuring segment performance, presenting live feedback on errors and deficits, and demonstrations of relevant expert-like behavior and performances.

Note: there is no reason for these activities to be in a particular ordering. For example, monitoring will happen both before and after recommendation. With these coaching activities defined, let us now look at the low-level components of the VC framework.

6.3 Framework

The VC consists of a performance library, a coaching progress manager, a task manager, a performance manager, and a score card. Briefly,

1. The performance library ($\mathcal{L}$) is a data base of relevant information from previously selected and logged task performance data.
2. The coaching progress manager (CPM) performs the `MONITOR` activity of the VC to update the coaching mode (m) as well as `RECOMMEND`s deliberate practice (DP) as needed.
3. The task manager (TM) configures a practice module based on information provided by CPM (m and DP), presents it to the trainee, and manages the real-time (*Online*) mode activities of the VC.
4. The performance manager (PM) reads the data (TD) logged and generated by the TM to `EVALUATE` the current performance, summarize errors and deficit metrics (`CRITIQUE`), and retrieve relevant demonstrations (*Demo*) from the performance library (`DEMONSTRATE`).
5. The score card (SC) presents the information generated by the PM ($Perf$ + $Demo$) to the trainee in a user-friendly interface.

Figure 6.2 outlines the VC framework flow when a skill is being learned and developed. The VC components are indicated using the sub-routine blocks and we will discuss them in more detail in the oncoming sections. Let us look at the flow from start

through end.

The trainee enters the flow on the left (*New / Existing Skill*) by choosing a skill that he is interested in learning and developing further. CPM suggests a coaching mode (m) based on previous history of the trainee's performance on this skill. TM presents a task module with the suggested coaching mode to the trainee. The trainee goes through the module and performs the task. Task data *TD* from the trainee's performance is processed by PM and the performance data *Perf* is stored on a data base under the trainee's performance history records. PM also extracts relevant demonstrations (*Demo*) for the trainee. *Demo* and *Perf* are sent to the score card (SC) to present them to the trainee. The trainee interacts with the SC information and eventually has two options – continue coaching or exit coaching (*Terminate Session* on top). CPM also generates and stores DP parameters onto a file if deliberate practice is to be recommended ($m = -1$). In case, CPM determines that the current skill has been developed up to an expert-level proficiency ($m = \text{Inf}$), the current skill coaching is marked as completed (*Completed Skill* on top) and the flow terminates. Upon continuing the coaching session, VC presents another repetition of the task module with updated coaching mode and the coaching cycle continues (marked by solid red →)

With this flow explained, we will present the internal flow and architecture of each of the VC components.

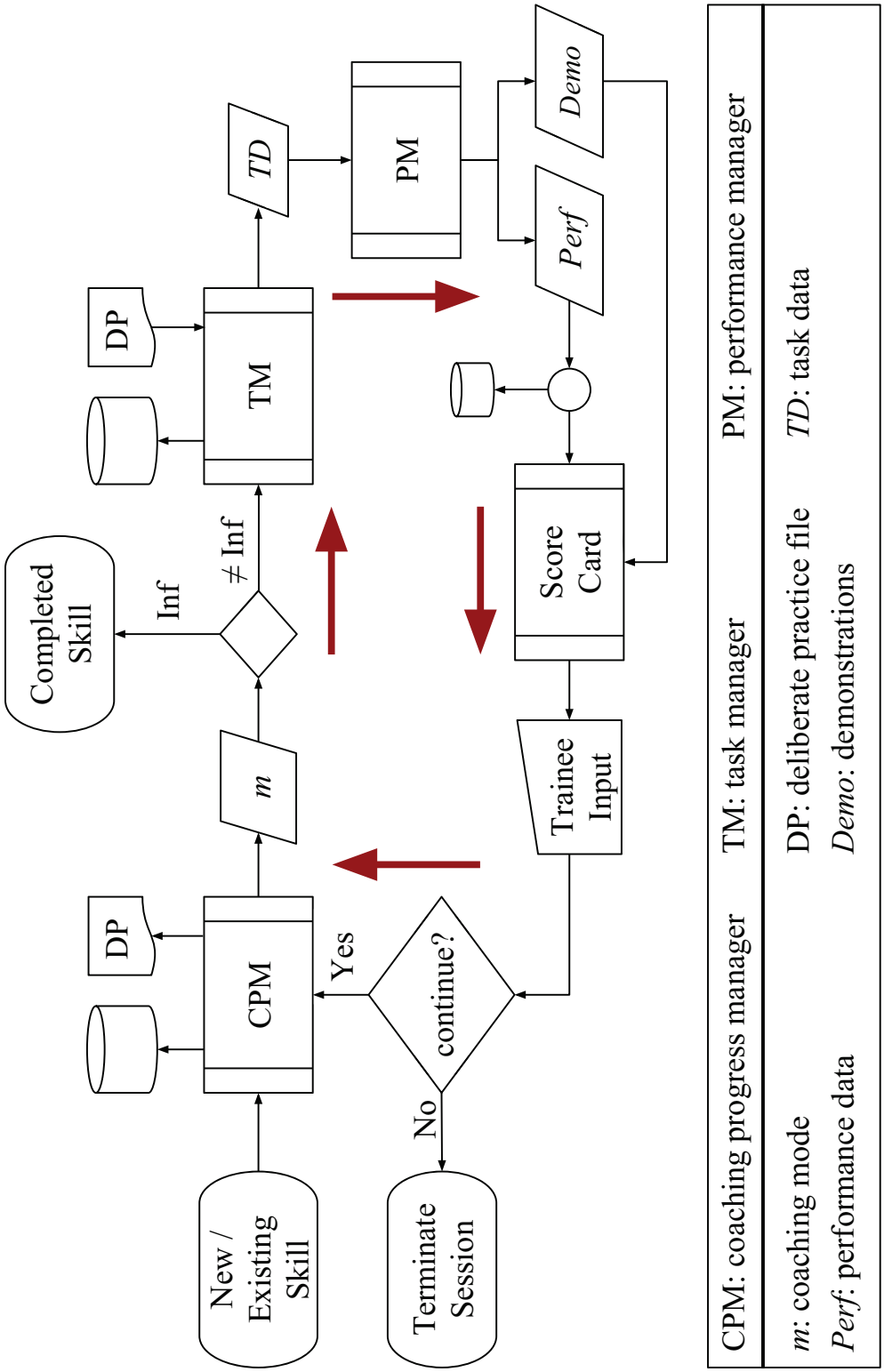


Figure 6.2: A flow chart explaining the information flow and components in the proposed VC

6.3.1 Performance Library

The VC relies on the existence of a rich data base of performances for all skills it provides coaching for. We refer to this data base as the 'performance library' denoted by the symbol $\mathcal{L}$. It is used by other VC components to perform each of the five activities based on real data from other users (trainees and experts). Here are some of the requisites for $\mathcal{L}$:

- must contain a uniform distribution of proficiency-level for each skill ranging from beginners to experts,
- must contain longitudinal samples tracking a user's proficiency history,
- the longitudinal data must contain users who had varying baseline proficiency,
- the performance data for each entry must contain task- and segment-level metrics based on instrument and console motion as well as errors and deficits,
- the performance data must contain a timestamped task progress log (indicating task segment-level flow),
- a subset of the performance data should contain a timestamped motion data log as well as the accompanying timestamped performance video.

Potentially, new trainees' data may be used to expand the library.

6.3.2 Coaching Progress Manager

The coaching progress manager (CPM) performs two of the coaching activities – `MONITOR` and `RECOMMEND`. CPM fetches all available trainee performance history to deliver on these activities. Figure 6.3 shows the flow diagram and internal components of CPM.

Monitor: Coaching Modes

The VC framework presents information to the trainee based on the current mode of coaching. We denote the coaching mode by m. This includes information displayed during the pre-, intra- and post- phases of the performance. The coaching mode is updated based on the trainee's proficiency history. The coaching modes are related to the four stages of learning developed by Noel Burch. Gawande[8] wrote about these in the context of surgical coaching as:

> Expertise, as the formula goes, requires going from unconscious incompetence to conscious incompetence to conscious competence and finally to unconscious competence.

Similarly, the coaching modes are related to various roles of a coach including a complete hands-on teacher, a mentor who intervenes as needed, and a hands-off guide. The different coaching modes are as follows:

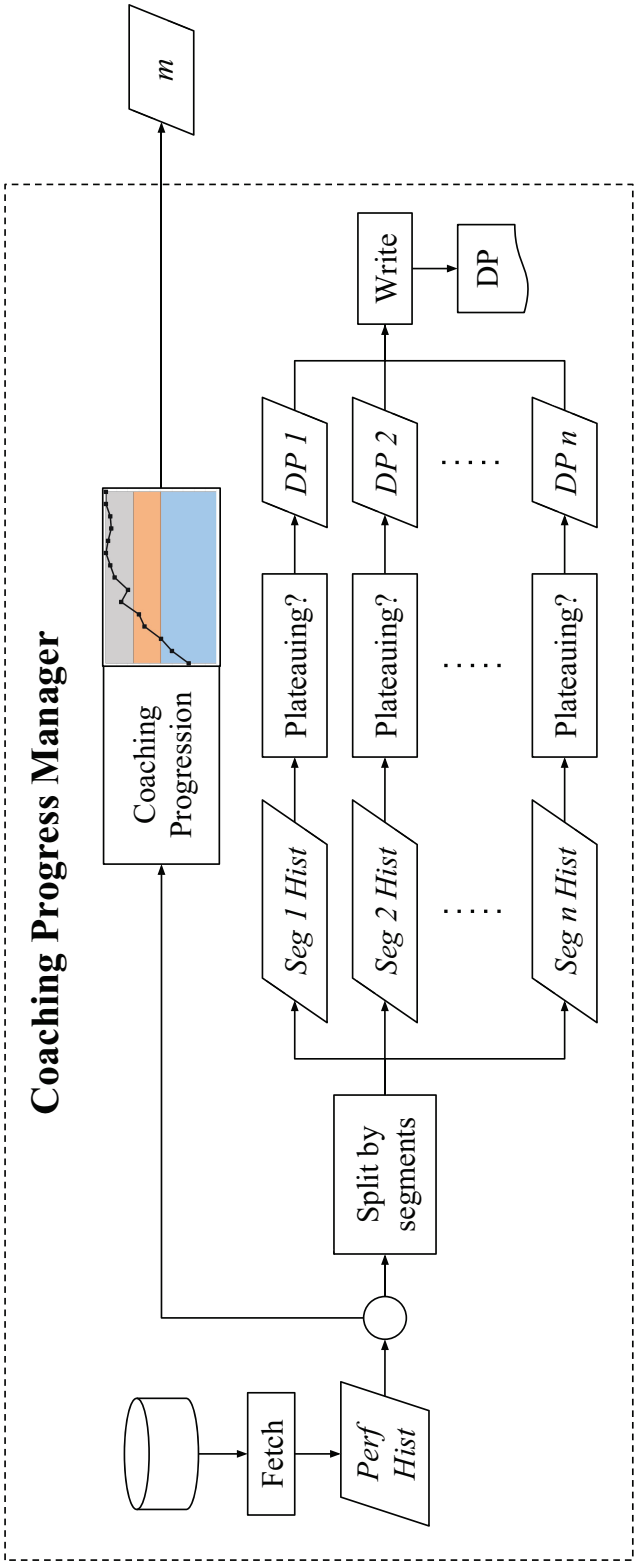


Figure 6.3: Coaching Progress Manager: flow chart diagram

BEGINNER ($m = 0$)

Any trainee (irrespective of their baseline proficiency) starts a new skill development in this mode. A complete didactic session is presented (`DEMONSTRATE`) to the trainee about the skill – errors that can occur, bad practices, complete expert demonstration. The goal of this mode is to orient the trainee to the task module and obtain task performances to estimate their baseline proficiency.

TEACH ($m = 1$)

This relates to the stage of unconscious incompetence in learning, while the VC plays the role of a hands-on teacher. The trainee is presented teaching cues for the skill during the performance to teach the ideal behavior and identify the errors and deficits that can occur. The VC also presents textual prompts indicating errors in performance and deviations from ideal performance. We presented such teaching cues in the context of robot-assisted needle passing skill in Chapter 5.

METRICS ($m = 2$)

This relates to the stage of conscious incompetence in learning, while the VC plays the role of an observer and mentor and intervenes when needed. The trainee is presented regular task modules to begin with (without any teaching cues in the form of graphical or textual overlays). The VC tracks live performance of the trainee in the background and intervenes by presenting relevant teaching cues whenever perfor-

mance deviates away from ideal beyond certain tolerance levels. The VC intervention disappears once the trainee is back to ideal behavior or the task segment is complete.

USER ($m = 3$)

This relates to the stage of conscious competence in learning, while the VC plays the role of a hands-off guide. The trainee has reached a proficiency level where they perform ideally provided they concentrate on performance details like errors and deficits and avoid them actively. The VC provides a help button for the trainee in case a quick reference to teaching cues is needed. However, unlike the METRICS mode, the teaching cues are presented only upon the trainee's request. The cues disappear upon trainee's request or once the current task segment is complete.

GRADUATE ($m = \mathbf{Inf}$)

This relates to the stage of unconscious competence in learning. The trainee no longer requires teaching of learning elements, errors and deficits. The VC provide no form of intervention and lets the trainee practice the skill in a regular fashion.

DELIBERATE PRACTICE ($m = -1$)

This relates to stagnant or halted learning. The trainee's proficiency is not yet at the unconscious competence stage of learning, but has stayed constant without signs of progress despite the use of teaching cues. The VC presents a sub module or variation of the regular task aimed at task segments which have not shown im-

provement. The trainee repeats in this mode until his/her proficiency shows signs of stable improvement. The deliberate practice mode can occur anywhere along the proficiency curve.

Note, a trainee's proficiency progress can vary. It is not necessary that it is non-decreasing in nature. Skill attrition is common at all levels of attained proficiency. A trainee can go from USER to METRICS or TEACH mode as well.

Figure 6.4 shows a cartoon example of coaching mode and proficiency progression. For this skill, the boundaries between TEACH, METRICS and USER modes are set at 50% and 75%. At BEGINNER mode, the baseline proficiency is detected at 25% and continues in the TEACH mode. By repetition 4, the trainee enters the METRICS mode of coaching and progress into the USER mode by repetition 6. The VC declares the trainee as GRADUATE at repetition 15 observing stable proficiency (100%).

Recommend: Deliberate Practice

In addition to coaching mode progression, the CPM identifies stagnated learning (plateauing) in the trainee's performance history to recommend focused DP sessions. The CPM analyzes performance data and detects plateauing of skill at the segment level. The information from all the segments is combined and stored on a file under the trainee's records. The file is used by TM to inject the focused task modules to the trainee.

In summary, the CPM fetches and analyzes trainee performance history and out-

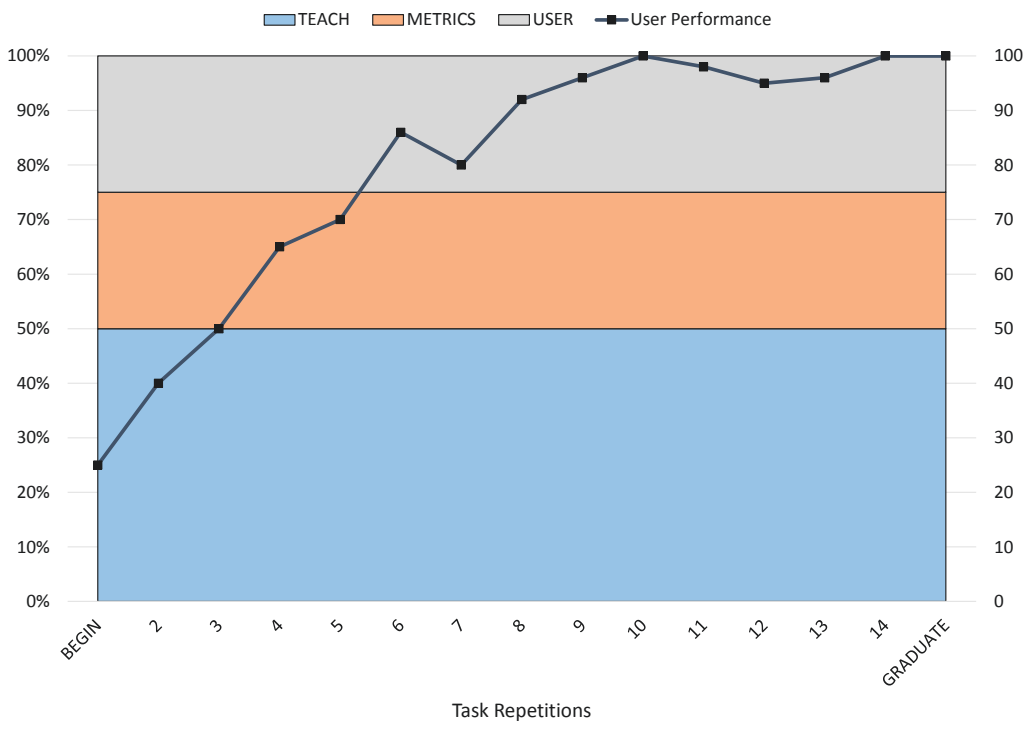


Figure 6.4: A cartoon sketch of coaching mode progression for a trainee

puts a coaching mode. A deliberate practice file is stored, if any.

6.3.3 Task Manager

Based on the current proficiency level (obtained from CPM), the VC adapts the current skills task via the task manager (TM). TM consists of a module generator, task progress manager (TPM), data logger and performance analyzer. Figure 6.5 shows the flow diagram for TM. Based on the incoming coaching mode, the module generator creates the regular task module or a customized DP task module. The module generator is a large framework by itself, and we will not go into details of how it should be developed. We have indicated the required inputs and outputs of such a component.

Task Progress Manager

The task progress manager maintains information about the task flow at the segment-level. All coaching activities rely on this information including online teaching interventions, segment-level performance evaluations and deliberate practice recommendations.

The task flow can be modeled as a directed graph as shown in Figure 6.6. The nodes in the graph represent state of the task S_x and edges are action sequences $A_{x,y}$ that should be performed to progress from one state to the next. The task state can be described by the tuple $S_x = (c_1, c_2, \ldots, c_n)$, where c_i contains context information

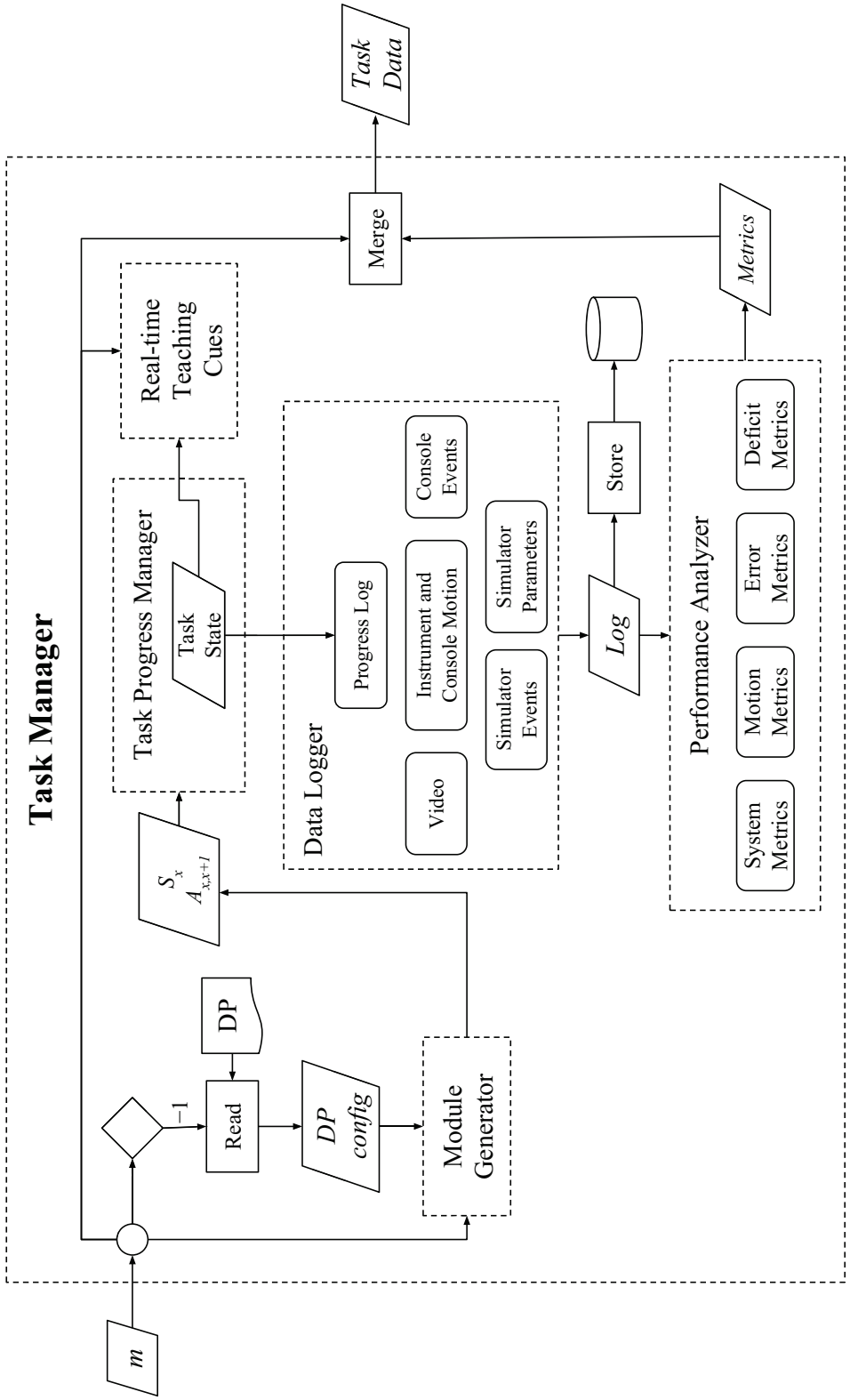


Figure 6.5: Task Manager: flow chart diagram

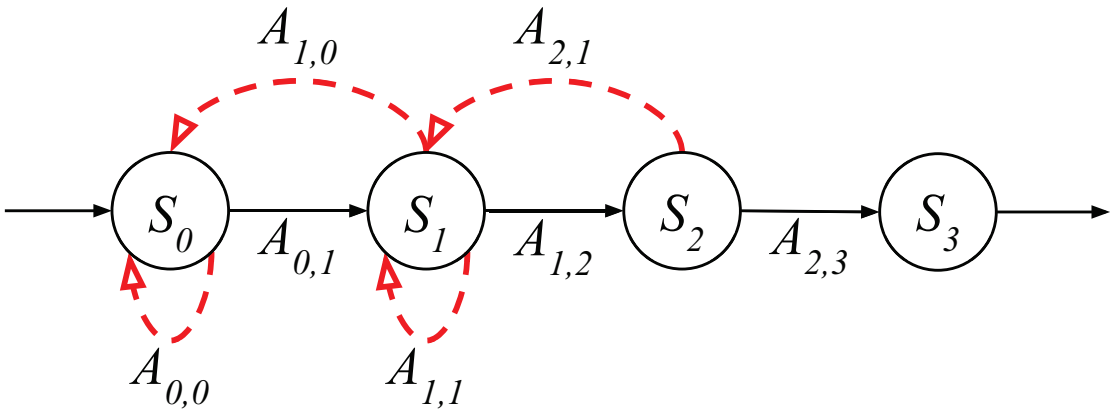


Figure 6.6: Task progress manager using a directed graph structure

about the task. This information may include the state of instruments, objects and targets as well as events. The actions, $A_{x,y} = (a_1, a_2, \ldots, a_q)$ are the interactions between instruments, objects and targets that occur in a state to induce a transition to the next state S_y. In addition, tasks are associated with a protocol that defines a set of rules that should be followed by the trainee for successful task completion. TPM relies on information about the task state at any given time and the task-specific protocol to monitor task progress. The module generator sends the task state S_x and action sequence $A_{x,y}$ information to TPM.

Data Logger

Different forms of data are available during the task execution in the VC. The data logger captures these data types along with timestamps so that TM can compute performance metrics on them. Additionally, the VC can provide self-review opportunities

to the trainee as well as use the task data log for expanding the performance library $\mathcal{L}$. The following data are logged:

- instrument and console motion,
- console user events,
- performance video,
- simulator events,
- simulator parameters, and
- task progress.

Performance Analyzer

This component of the TM measures the trainee's performance in an online (real-time) and offline (post-completion) setting using a number of metrics established in literature for assessing skill in manipulative and dexterous tasks. The performance analyzer uses data from the logger to compute these metrics at task- and segment-level (using the task progress log). The following type of performance metrics are computed:

- System Metrics: based on system events and use like number of times clutch was pressed and so on.
- Motion Metrics: based on motion efficiency in instrument usage and console workspace usage like instrument path length, ribbon area, number of movements

and so on.

- Error Metrics: based on errors committed by trainee like dropping an object, breaking the suture, applying cautery at wrong target and so on.
- Deficit Metrics: based on learning elements like deviation from ideal grasp orientation while needle passing and so on.

In summary, the TM takes the coaching mode as input and generates a task module, logs task execution data and computes performance metrics. The output of the TM is a combination of performance metrics and the coaching mode (m) and is referred to as *Task Data (TD)*.

6.3.4 Performance Manager

Post completion of a task module, the trainee is presented an evaluation and feedback about their performance. The performance manager (PM) undertakes the remainder of three activities of the VC *viz.* `EVALUATE`, `CRITIQUE` and `DEMONSTRATE`.

Figure 6.7 shows the flow diagram of the PM. Essentially, the task data (TD) from TM is analyzed using the pairwise comparisons framework for objective skill assessment described in Chapter 4. This generates skill scores at task and segment level for the trainee's performance. PM uses this score information along with the metrics to determine segments that require feedback. PM also retrieves expert performances from $\mathcal{L}$ for the relevant segments as demonstrations ($Demo$) for the trainee to look

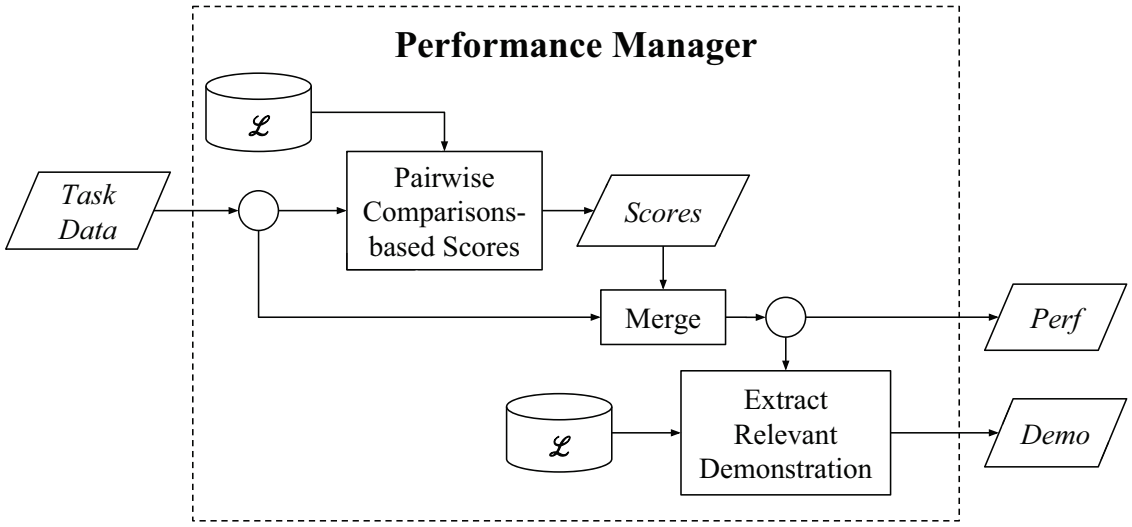


Figure 6.7: Performance Manager: flow chart diagram

at. The combined performance *Perf* (performance scores + metrics) and *Demo* are sent to the score card component.

6.3.5 Score Card

This is the user facing component of the VC framework. The SC lays out relevant information generated by all other components (CPM, TM and PM) for the trainee's review. We will not go into details of the SC architecture since that will involve explanation of other user interface (UI) components as well. Instead, we show a concept sketch of SC in Figure 6.8. At the very top, the current skill and user name are displayed. Following this, a timeline of the task execution with the constituent segments is shown. The trainee may click on these segments to view a segment-level version of the evaluation, feedback and demonstration. The current view shows the

task-level information. Underneath the timeline, a tab-based panel shows information about performance scores, errors, deficits, demonstrations and deliberate practice (if any). Instead of score values (which are visible upon hovering) a ranking is shown for each segment and the overall task using a gradient widget. This gives a quick summary to the trainee about *how did I do?* and *where was I wrong?*. Red indicates beginner level performance, while green indicates expertise proficiency. The `ERRORS` and `DEFICITS` tabs show the task-level error statistics and learning element deficits (*why was I wrong?*). Segment-level data is presented if the user clicks on the relevant segment in the timeline. The `DEMONSTRATIONS` tab shows the *Demo* data received from PM (*how do I do it correctly?*). Finally, the `DELIBERATE PRACTICE` tab lists possible DP modules recommended for focused skill development (*what do I do to improve?*). A performance video and the trainee's learning curve are displayed in the bottom right corner as well.

This concludes the framework description of the VC. We will now introduce our current implementation approach for realizing the automated virtual coach.

6.4 Implementation

We will describe preliminary implementation details of the VC on the da Vinci® Skills Simulator™ .

Figure 6.8: Score Card: a concept sketch. The content from the VC components is displayed to the trainee post completion of a task execution for review, feedback and self-reflection.

6.4.1 Software Architecture

We obtained proprietary code and software from Intuitive Surgical Inc. (ISI) and SenseGraphics AB (SG) by establishing a non-disclosure agreement between the three entities (including Johns Hopkins University (JHU)). The software framework contains the simulation infrastructure for some of the training task modules available on the commercial product. We will refer to this software package as Simulation Sandbox (SS) – a sandbox for testing surgical training and coaching interventions in a VR simulation environment.

H3DAPI

The SS is based on the open source library H3DAPI.[213] This library is developed and maintained by SG. The H3DAPI is an open source haptics and graphics software package that uses the OpenGL library for graphics rendering and HAPI library for haptics rendering. It also provides wrappers for standard physics engines like Open Dynamics Engine, Bullet, and PhysX. The platform provides three programming interfaces – X3D (an ISO standard XML-based format for representing 3D graphics), Python and C++.

The software architecture is modular and divided into *Nodes* and *Fields*. Fields are data containers and can be connected to each other in a directed graph for sending events and passing data. These directed connections are referred to as *routes*. On the other hand, nodes are containers for fields. Nodes provide a better management and

form the building blocks of the scene graph. For example, `Material` can be a node for describing material properties for the geometry in the scene. The material node can contain fields corresponding to `diffuse color`, `shininess`, `transparency` and so on. Like a graph, a node can be contained within another node as its field. In the current example, the material node can be a field contained within the `Appearance` node that specifies the visual properties including material, textures, shadows, etc. This architecture allows for simple data passing, event handling, and state updating.

While this software library and architecture is available as an open source project, other components of the SS are proprietary to ISI and SG respectively. For example, simulation models and parameters for the da Vinci instruments are proprietary in nature. Similarly, robot-assisted surgical training specific developments of Nodes and Fields using H3DAPI by SG are proprietary in nature as well.

6.4.2 VC Task Manager

We implemented the task manager along with the three coaching modes (TEACH, METRICS and USER) of the VC using the node-field architecture of SS on the needle passing module available. The real-time teaching cues and the corresponding randomized controlled trial (RCT) presented in Chapter 5 was conducted using this implementation. Owing to the existing NDA we cannot go into further details of the implementation. However, the RCT was an excellent demonstration of the feasibility of development and validation of concepts and components described earlier

within our proposed VC framework. Currently, we are implementing the performance manager components of the VC using concepts described in Chapter 4.

6.5 Summary

In this chapter, we laid out the motivation and need for automated surgical coaching by giving examples from literature that have shown and proven the value of coaching in surgical skill acquisition, development and retention, and listing current limitations with regards to time, cost and scalability of such manual coaching interventions. We have presented the first end-to-end framework, virtual coach (VC), to provide automated surgical coaching for technical skills development. We listed the five core coaching activities that the VC delivers viz. demonstrates expert behavior, evaluates task and segment performance, critiques errors and deficits, recommends deliberate practice and monitors skill progress. We presented the core components of the VC – performance library, coaching progress manager, task manager, performance manager, and score card. We gave a glimpse into current implementation and software architecture using H3DAPI and the Simulation Sandbox.

Currently, we are setting up a data collection protocol for building the performance library $\mathcal{L}$ starting at our institution first. We then, plan to conduct an initial face validation study and a follow-up construct and content validity study by recruiting medical students, surgical residents and faculty surgeons.

Simulation training and coaching have shown success in many fields outside of surgery. Surgical educators have advocated for the use of simulation and coaching as tools to provide training outside the OR and enabling lifelong learning. Our proposed VC delivers on those two recommendations and has shown initial success in delivering automated coaching in VR simulation training.

Chapter 7

Discussion and Conclusion

In this thesis, we have presented surgical data modeling techniques that are building blocks of an automated virtual coaching framework (VC) to deliver directed and individualized learning supplemented by critical feedback and relevant demonstrations.

Over the last two to three decades, surgical education and patient care policies have motivated research towards resident training, lifelong expert learning and simulation-based technical skills development. Global rating scores and checklists based assessment, immediate and concurrent expert feedback and video debriefing based coaching emerged as widely studied and validated solutions addressing these themes of research. Restricted duty hours, expert surgeon time and cost effectiveness have deterred large scale adoption of these manual approaches. Automated approaches modeling video, motion and sensor data have shown initial success at skill

assessment, workflow analysis and video indexing and retrieval. Lack of fine grain evaluation and critique, directed learning and expert demonstrations have restricted these methods from entering surgical training labs and OR. No known framework exists to deliver expert-like coaching in an automated and individualized setting. Such a framework should be able to model task flow, extract activity context, evaluate fine-grain skill, detect errors and deficits, recommend deliberate practice and demonstrate ideal behavior. Our proposed VC framework has been designed and developed to perform these essential coaching activities that are missing in literature and to address the above limitations of manual coaching.

7.1 Summary

Chapter 1

In this chapter, we presented a dialogue about current surgical education, coaching in other fields, instrumentation of the OR and surgical data modeling, success of crowdsourcing, value proposition of virtual reality (VR) training and lack of coaching in current VR trainers. We laid out the thesis statement and contributions along with a short summary of topics discussed in each chapter.

Chapter 2

In this chapter, we covered concepts and terminology used in the remainder of the thesis. We described the commercial da Vinci® Surgical System and da Vinci® Skills Simulator™ from Intuitive Surgical Inc. We presented our data recorder system for collecting intra-operative surgical performance data from the OR and training labs. We listed currently validated motion metrics for surgical performance evaluation. We introduced concepts from the crowdsourcing domain and Amazon Mechnical Turk™

Chapter 3

In this chapter, we described current state and our work in the domain of surgical activity modeling. We presented a pipeline for surgical phase recognition in OR procedures by summarizing shorter time segments, assigning posterior scores to them and generating a temporal phase labeling by running a segmental inference over the posteriors. We presented two approaches for summarizations using a robot-assisted laparoscopic hysterectomy data set. First, we showed phase recognition accuracy of 75% using system events-based summarization features like cautery (energy) usage and instrument identification to perform phase recognition. Second, we ventured towards surgical context-based summarizations. We performed a pilot study to show that crowdsourcing can generate reliable and valid context summarizations containing information about surgical activity being performed, instrument use and purpose, occurrence of events like bleeding and irrigation, and presence of objects like gauze

and needle. We showed initial validation of surgical phase recognition using these crowdsourced summarizations.

Chapter 4

In this chapter, we described current state and our work in the domain of automated and objective surgical skill assessment. We presented a pipeline for fine grain evaluation of skill by representing segment-level performances by motion metrics, comparing such performance representations in a pairwise approach (*preferences*), generating a ranking-based percentile score using such preferences and regressing an overall task-level score based on such percentile scores. We obtained reference annotations for preferences through crowdsourcing and presented extensive validation of reliability and validity of such preference ratings. We showed that a classification algorithm (preference classifier) can be trained to predict the preferences accurately. We demonstrated equivalence in crowd and expert preference ratings by comparing the outcomes of preference classifiers in terms of accuracy, segment scores and task scores. We validated the preference classifiers across data sets with consistent accuracy in predicting preferences for bench-top interrupted suturing and virtual reality needle passing tasks. We presented a reliability analysis of absolute versus relative ratings for segment-level skill assessment.

Chapter 5

In this chapter, we described current state and our work in the domain of feedback and demonstration for surgical training. We introduced the concept of learning elements of a surgical skill as key feedback and demonstration points for effective training. We introduced the concept of deficit-based metrics targeted at measuring deviations from ideal performance of these learning elements. We introduced the concept of teaching cues as a tool to demonstrate ideal performance of the learning elements. We listed and described learning elements, errors and deficit metrics and real-time teaching cues in the context of a needle passing task in a VR simulation setting. We presented a pilot randomized controlled trial to study the effectiveness of these real-time teaching cues in learning a needle passing task on the da Vinci Skills Simulator. We observed significant improvement in one of the learning elements between the control and experiment populations. We received positive comments about the teaching framework and the need for one to learn such skills from the study population.

Chapter 6

In this chapter, we described current state and limitations of manual surgical coaching and proposed our automated virtual coaching framework (VC). We introduced the key coaching activities of the VC *viz.* demonstrate, evaluate, critique, recommend and monitor and their significance in providing individualized and directed

learning. We presented the main components of the VC *viz.* performance library, coaching progress manager, task manager, performance manager and score card. We described the learning of a new skill and the coaching process using a flow chart diagram. We presented key underlying concepts of the VC components like coaching modes and task progress manager using flow diagrams and concept sketches. We described the H3DAPI library and showed the feasibility of realizing the VC using it.

7.2 Limitations and Future Work

Surgical Activity Modeling

We investigated surgical phase analysis using system events data, however, phase information is distributed across different forms of data - video, motion and system events. Future work should look at combining multiple modalities to capture complementary information about surgical phases. Our experiments were limited to a single surgical procedure data set and further investigations must be performed to verify the validity of our hypothesis and its scaling. Additional limitation of the data set was its size and variability. We must scale up the data set so there are a sufficient number of samples belonging to different sets of parameters like operating surgeon, patient's anatomy, for statistically significant analysis and results. Future research must consider this when generating new data sets. The crowdsourced context summarizations were obtained on a subset of the data set and future work should validate

the findings on a larger data set and on other surgical procedures.

Surgical Skill Assessment

One limitation of our work on segment-level skill scoring is the fact that our approach requires prior segmentation of the task into constituent segments. This assumes both that such constituent segments exist and that the resources or infrastructure to perform this segmentation exist. While such segmentation can be obtained freely in the VR environment, validation using current approaches on bench-top data should be performed in future. An interesting and open question for future work is whether such segment-level assessment is more effective in skill development compared to traditional global scores. Similarly, it is not yet established whether the effectiveness of the framework is sensitive to the granularity of analysis and whether assessment at levels finer than maneuvers in the task, such as gestures, may be important for surgical skills acquisition. Current results are based on simple motion metrics and future work should look at error and deficit metrics as well as video-based features for improving the accuracy of the framework. While we presented a reliability analysis comparing absolute and relative crowdsourced ratings, the final outcome of the pipeline is automated task and segment scores. Future work should compare automated assessments using absolute and relative ratings. The percentile scoring for segments is a simple ranking approach. However, researchers have developed better ranking algorithms in the domain of video gaming, sports, recommendation systems

like Elo rating, TrueSkill™ and so on. Future work should explore these techniques and compare them for prediction accuracy in task scores. Finally, the effectiveness of the framework relies on the availability of a rich performance library which does not exist currently. Future data collection at institute, state and country levels should be planned to collect a representative corpus of performances.

Feedback and Teaching

Our current work was narrow and specific to VR simulation based needle passing, however, it was the first attempt in doing so in an automated manner. Future work to extend the framework to other surgical skills should be performed to test whether learning elements and deficits can be described in the context of others skills. We performed a pilot study using engineers and non-surgical trainers to show feasibility and validity. However, true effectiveness can be proven only once residents and expert surgeons are included in the study sample. A limitation of the current work was the longitudinality of the experiments. Future work should account for learning, and design experiments with larger number of training sessions. Our results are currently based on performance metrics - not all of which are validated at a large scale. Future work should include expert-assigned GRS for confirming the effectiveness in learning observed by the performance metrics. While we mentioned and introduced the concept of deficit metrics in a task-specific manner, perhaps some deficits are more generic and applicable across skills. It is an interesting, new and open area

of research in the domain of surgical training to find underlying signatures common across different skill sets.

Automated Virtual Coaching

Our VC framework is in the nascent stage of development. Most of concepts are in theory and have not yet been implemented on an actual system. A large array of future studies should be performed to answer various questions and testing the different coaching activities. First, effectiveness of providing a segment-level scoring should be tested. Second, effectiveness of presenting errors and deficits should be tested – both in an online (real-time) and offline (post-completion) setting. Third, effectiveness of online coaching (real-time teaching cues) among a surgical audience should be tested. Fourth, effectiveness of the different coaching modes as well as coaching mode progression should be tested along with its validation. Fifth, effectiveness of recommending deliberate practice sessions and the validity of the framework in generating those should be tested. Sixth, effectiveness of the score card i.e. a combination of the different coaching activities along with its usability should be tested. Finally, a comparison of no coaching versus manual coaching versus automated coaching by our VC should be performed. Another open research question is that of acquiring a performance library that can be useful for delivering valid and reliable coaching.

7.3 Remarks

The value proposition of simulation-based training has been a subject matter for the past two decades now.[29,40] With the initial results presented in this thesis, we have presented a solution towards addressing some of the limiting factors of large scale adoption and use of VR training.

We also note that technical skill is one component of the overall performance in the operating room, and further work to incorporate pre-operative and post-operative skills can help predict patient outcomes. There is however, wide variation in technical skill among practicing surgeons,[121] and poor technical skill is associated with an increased risk of adverse patient outcomes including death and re-operation.[121,214,215] We believe that solutions like our VC, for providing effective training outside the OR are needed to resolve the ethical concerns about training on the patients. We believe that such improvements in performance with automated surgical coaching will transfer without attrition to the operating room, and eventually affect safety and quality of patient care.

Surgical data sets are limited in size and variability. Most of the data sets face challenges of privacy and legal concerns from the patients and hospital administrations. Existence of large data sets is crucial for the development of effective automated tools for efficient surgical interventions like coaching. The problem of obtaining such large data sets at institute- , county-, state-, national- and international-levels is a challenging one. Future steps as a community should be taken towards the building

of a larger and diverse data corpus.

Appendix A

Data Sets

A.1 MultiSite Suturing Data Set

As the name suggests, the *MultiSite Suturing* data set was collected across multiple teaching hospitals as the study sites. However, only a subset of the data is currently being used. This data belongs to just one of the study sites.

Goal

The goal of the user study was to develop objective metrics for skill assessment using instrument motion data obtained from the da Vinci API across multiple medical institutions.[54, 59, 171]

Task Description

A bench-top simulation model (shown in Figure A.2) was used to setup the task of interrupted suturing. The simulation task model is commercially manufactured by The Chamberlain Group (SKU: 4026; `https://www.thecgroup.com/product/robotic-skin-suturing-pod-4026/`). We used the linear defect of the model and oriented the task pod to setup a vertical defect repair. Three sets of entry and exit targets were marked using a colored pen on the tissue model. 3–0 Vicryl sutures were trimmed to a 10cm length for the study.

Study Participants

The subset of data described and used in this thesis was obtained from 14 surgical residents (novice to the dVSS and had no experience in robot-assisted suturing) and 4 attending surgeons (experts with prior experience on using the dVSS and performing robot-assisted suturing).

Study Protocol

The participants were asked to perform three repetitions of the interrupted suturing task without any breaks in between the repetitions. The study was designed to obtain a longitudinal data set from each of the participants. In case of the residents (novices), there was a gap of at least a week between two study sessions. In case of the attending surgeons (experts), two sessions of such three repetitions were collected

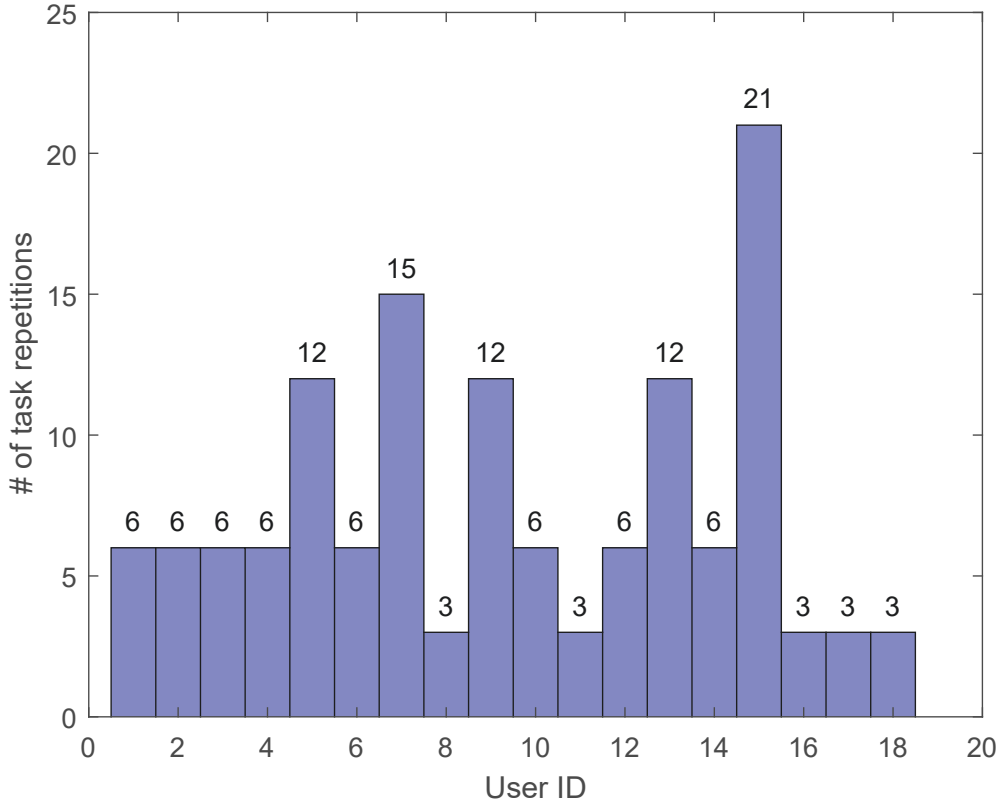


Figure A.1: MultiSite Suturing: number of trials per user

in the same visit with a gap of 10 minutes in between the two sessions. A distribution of the number of task repetitions by each user is shown in Figure A.1.

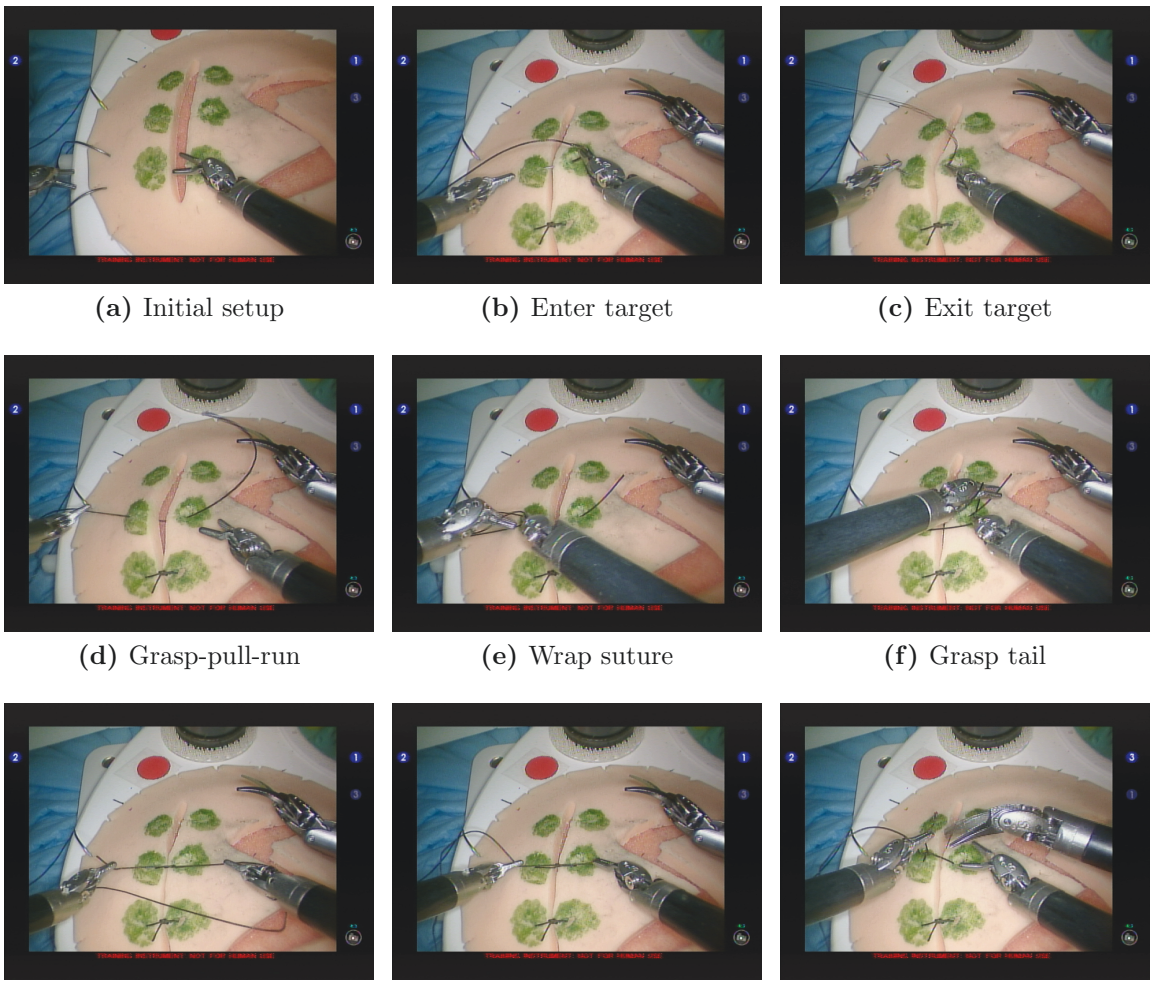
The task protocol was a bit loosely defined. There were no specific instructions provided to participants about the suturing technique. They could choose the number of throws to secure the knot. At times, participants used left over suture from a previous performance as well. The participant was free to perform a left-handed or right-handed suturing irrespective of their dominant hand.

Hardware

The dVSS S model was used at the study site from which the subset data is used. Two large needle driver instruments were mounted on the primary PSMs to manipulate the tissue and needle. A curved scissors instrument was mounted on the third PSM to cut the extra length of suture after completing the interrupted suturing attempt.

Task Flow

Typically, the task execution started by passing the needle from one side of the tissue and driving it through the other side of the repair while aiming for the marked targets. The needle passing could be broken down in two drives – passing the needle through the entry side target and pulling it out through the center, and then passing it from the center through the exit target and pulling it out from the exit side. This was followed by a sequence of grasp-pull-run actions to pull the suture out from the other side leaving a tail on the entry side. Holding the needle end, a wrapping motion is performed to put one or two loops of suture on the other instrument. The instrument with the wraps is used to grasp the suture tail on the entry side. The tail end is pulled through the wraps resulting in the first throw of the knot. This is followed by a sequence of multiple throws (ranging from one to five) to finish the knot tying. Afterwards, the scissors instrument on the third arm of the dVSS was used to cut the extra length of suture remaining. Figure A.2 shows images from a task execution for



(a) Initial setup **(b)** Enter target **(c)** Exit target

(d) Grasp-pull-run **(e)** Wrap suture **(f)** Grasp tail

(g) Pull tail & complete throw **(h)** Subsequent throw(s) **(i)** Cut extra suture

Figure A.2: MultiSite interrupted suturing task execution

each of these steps.

Data Collected

A previously existing data archival system was used for this data collection.[54] The kinematics data contains all the fields mentioned in Table 2.1 except for the joint torques (which were added in a later version of the da Vinci API). The stereo video

was captured at a 30 Hz framerate and 720 (640 in some cases) x 480 resolution. Compared to the da Vinci data recorder tool described in Section 2.1.4, the older system did not reliably capture the da Vinci API events. A video framestamp was used to synchronize the kinematics and video frames, which meant that multiple kinematics frames had the same framestamp number. The video was collected using an analog capture card with a higher compression value resulting in poor image quality compared to the new data recorder tool.

Data Annotated

We annotated the data set for skill and activity segments.

OSATS Ratings

An experienced surgeon, watched a combined video for all three task repetitions from a session. They were blinded to the identity of the study participant while doing so. For each session, the rater used a modified OSATS form and assigned a score for each of the components. The skill component on *Use of assistants* was dropped as it was not relevant in our case. We extrapolated the session-level scores to each repetition in the session.

Activity Labels

A vocabulary for *gestures* and *maneuvers* was developed in consultation with surgical educators from our collaboration. Detailed description of the vocabulary can be found in Vedula et al.[104] We broke down the task into the following five maneuver categories to account for variability in how different participants performed the study task:

- ST1 – suture throw performed in two steps; passing the needle separately through each side of the incision or repair;
- ST2 – suture throw performed in one step; passing the needle through both sides of the incision or repair in a single motion;
- GPR – running suture out of tissue following a suture throw;
- KT1 – the first knot;
- KT2 – any knot thrown after the first knot.

In addition to maneuver categories listed above, our vocabulary for maneuvers in the study task included inter-maneuver segments (IMS). IMS represent portions of the task wherein participants performed certain actions in preparation for the next maneuver.

Two individuals, independent of each other, manually annotated video recordings of the trials for start and end of maneuvers within the task. A few labeled task repetitions are shown in Figure A.3.

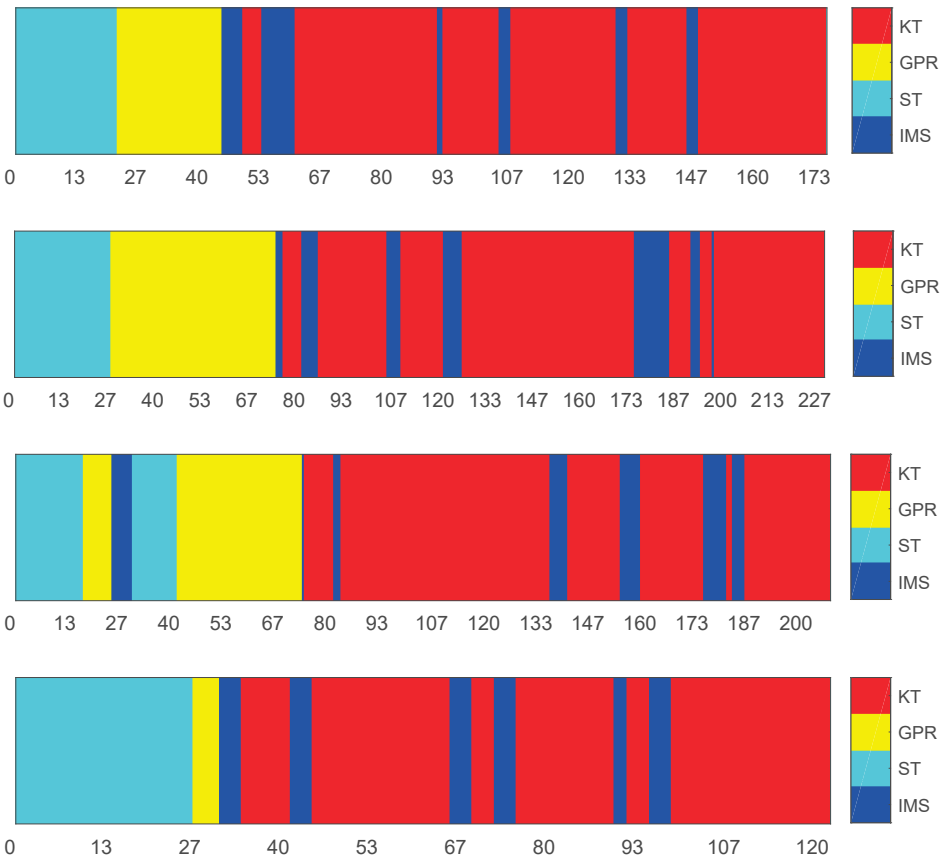


Figure A.3: MultiSite Suturing: maneuver flow samples from data set

Table A.1: MultiSite Suturing: maneuver distribution

Maneuver	Count
ST1	60
ST2	104
GPR	154
KT1	135
KT2	203
Total	1008

remaining maneuvers are either IMS or incomplete instances

Size

A total of 45 sessions were recorded, each consisting of 3 repetitions of the task, resulting in a data set of 135 interrupted suturing performances. A total of 1008 maneuvers were annotated.

Access / Restrictions

The data set is currently under a non-disclosure agreement with Intuitive Surgical Inc. and will need further processing to strip proprietary content out of the data before it is ready for public release.

Limitations / Pros

Currently published literature (as of December 2016) does not mention of a data set of this magnitude containing instrument kinematics and endoscope video. The only other data set with similar content is the JIGSAWS published in 2014 by our

group.[216] In comparison to JIGAWS, this data set is much more in the wild and realistic. The study participants in JIGSAWS were restricted to a fixed setup for the dVSS and were not allowed to use the camera and clutch pedals while performing the task. There was no such restriction during the MutliSite study. One of the major limitations of the MultiSite data set is the loosely defined task protocol that resulted in multiple variations of the task flow as described above. Also, the OSATS scores for the sessions were obtained from a single expert blinded to the participant identification. And, the OSATS scores were assigned for the entire session of three task repetitions and not for individual task attempts. This can, however, be resolved by conducting a new round of OSATS or GEARS ratings at the task repetition level with multiple reviewers.

A.2 ISI-SG-Sim Needle Passing Data Set

This is a virtual reality simulation-based training task data set that was collected at Intuitive Surgical Inc. (Sunnyvale, California).

Goal

The goal of this user study was to validate the effectiveness of a real-time teaching framework for surgical training using a VR simulator in a randomized controlled trial design.

Ethical Review

This study was approved by Western IRB (protocol #20121049) and conducted at Intuitive Surgical Inc., Sunnyvale, CA.

Task Description

A needle passing task from the dVSim was used for the study. This VR simulation has been developed by Intuitive Surgical Inc. and SenseGraphics AB. The task model consists of a 3D deformable tissue model. Eight pairs of entry-exit targets, circular in shape, are marked on the top surface of the tissue model. The target pairs were arranged around two concentric circles. A surgical curved needle is placed on the top surface at the start of the simulation as well.

Hardware

The data was collected on the dVSim Xi model located at Intuitive Surgical Inc.(Note: the Xi model SSC is similar to the Si model SSC for most of the user features).

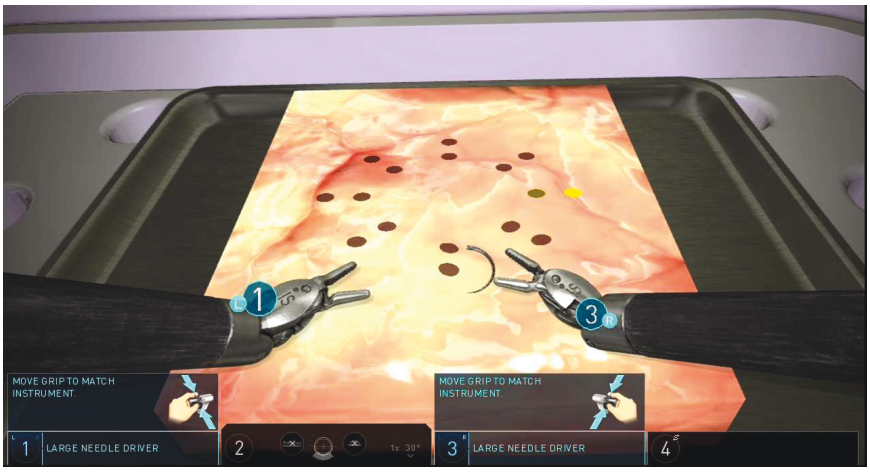
Task Flow

Owing to VR simulation, the task began in the same setup every time. Two virtual large needle driver instruments were provided on the primary PSMs to manipulate the needle. The simulation highlights the current pair of targets with yellow color. The

entry side target is indicated by a flashing yellow color. At the start, the participant picks up the needle located on the top surface of the tissue model. They pass the needle through current entry target and drive it to pierce through the corresponding exit target. Following this, they pull out the needle from the exit target. Upon successful completion, the yellow highlights move to the next set of targets and the task continues. Incorrect insertion missing the target circle requires a re-insertion and the highlights do not progress to next set of targets until current needle passing is successful. The task is complete at the successful completion of the eighth needle pass. Figure A.4 shows a sequence of images demonstrating the described task flow.

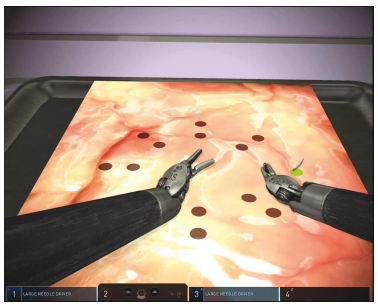
Study Participants

Study recruitment was performed within the Intuitive Surgical Inc. employee pool. An advertisement email was sent out inviting the engineering and clinical design and training groups. A total of 32 participants were scheduled for the study, of which two had to leave the study without completing the protocol. Of the remaining, six participants were clinical trainers at the training facility at Intuitive Surgical Inc. and have a good experience using the system and the simulator. Other 24 participants were from the different engineering divisions with varying levels of experience using the dVSS, dVSim and robot-assisted needle passing.



(a) Initial setup

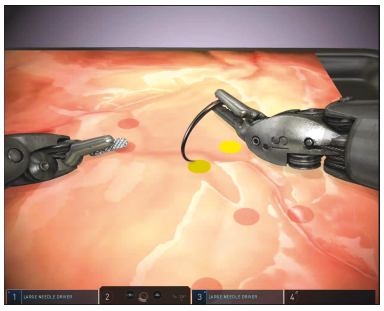
(b) Enter target



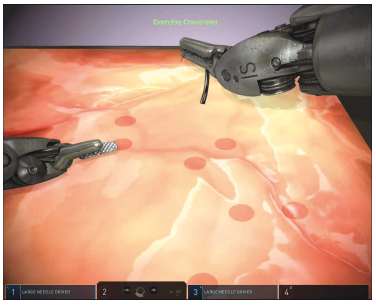
(c) Drive through

(d) Exit target

(e) Next target



(f) Last (eighth) target



(g) Task complete

Figure A.4: ISI-SG-Sim needle passing task execution

Study Protocol

The study was designed as a randomized controlled trial with stratification for the trainer and engineer groups. The control group performed the 'regular' version of the needle passing task (as described above), while the experiment group were presented real-time teaching cues during the needle passing task. The participant was randomly assigned to control or experiment group upon arrival. All participants filled a pre-study questionnaire about prior robot-assisted system usage experience. Both groups performed a baseline performance on the regular version of the task. This was followed by a sequence of three repetitions of the task for learning and developing the participant's skill. Control group practiced on the regular version all three times in a self-learning (independent) setting. Experiment group were presented visual overlays as teaching cues showing the ideal (expert-like) approach to needle passing. Different teaching modes were presented during each of the three repetitions. We explain these modes and the teaching cues in detail in Section 5.2.3. Afterwards, both groups performed a final test repetition on the regular version of the task. To complete their study participation, participants were asked to fill a post-study questionnaire about the task difficulty level and self-evaluation of learning. Experiment group answered additional questions about the clarity and quality of teaching cues, perceived effectiveness of teaching cues and modes. While, the control group was asked if they felt the need for external teaching and what teaching mode would have they preferred.

Data Collected

Under a mutual non-disclosure agreement between Intuitive Surgical Inc., Johns Hopkins University and SenseGraphics AB, we could write a data logger code similar in nature to the da Vinci recorder tool (Section 2.1.4) to collect MTM kinematics, virtual PSM and ECM kinematics and system events. In addition, we could capture other simulation parameters and objects as well as simulation events. For example, we recorded the curvature of needle, location of the targets and tissue, needle pose in 3D, and events like an instrument grasped the needle. A task progress log was recorded logging the different steps of the task flow like needle insertion and needle pull out. We recorded the left camera video from the virtual endoscope instead of a stereo video. A universal timestamp was recorded with each of the kinematics, events, task progress and video frames for synchronization. Study related questionnaire responses were collected as well.

Data Annotated

At present, no annotation has been performed on the data set. Maneuver segments can be computed automatically using the task progress log recorded as part of the data collection.

Size

A total of 102 completed repetitions of the regular version of the task were recorded. There were recordings for 42 repetitions of tasks with teaching cues shown to the participant.

Access / Restrictions

The data set was collected under two NDAs between JHU and ISI, and JHU and SG. It was collected under a Western IRB protocol with an ISI principal investigator. The final decision about release of data set will be with ISI.

Limitations / Pros

Previous studies have collected data from VR simulators including video and performance evaluation metrics based on kinematics, system and simulation events. However, none of the previous data sets have captured the raw kinematics signal which is unique in the ISI-SG-Sim data set. Such access to the raw kinematics allows for development of new assessment methods. Another unique feature of this data set is the log containing the task progress which enables automated task segmentation which is of value in generating targeted and individualized feedback. Finally, the simulation events and kinematics are also unique and have been captured for the first time. The ability to localize the needle (and other objects) has a lot of value for training and assessing skill at critical elements of needle passing. However, the data set does have

limitations. The simulation training task was work in progress during the study. This resulted in system crashes due to existing bugs in the simulation framework. This led to unpleasing moments with study participants and loss of data as well. The data set is not longitudinal enough to see learning effects and a future study should attempt a larger number of task repetitions and multiple training sessions. The study participants were non-surgeons and so the value of surgical training cannot be truly tested until a future study in a hospital setting is conducted.

A.3 WarmUp Hysterectomy Data Set

This data set was collected as part of a study with a VR simulator component as well as an OR procedure component. The *WarmUp Hysterectomy* data set is the data from the OR.

Goal

The goal of the user study was to test whether pre-operative warm-up using a VR simulator has effect on resident's intra-operative performance during a robot-assisted laparoscopic hysterectomy (RALH) procedure.[181]

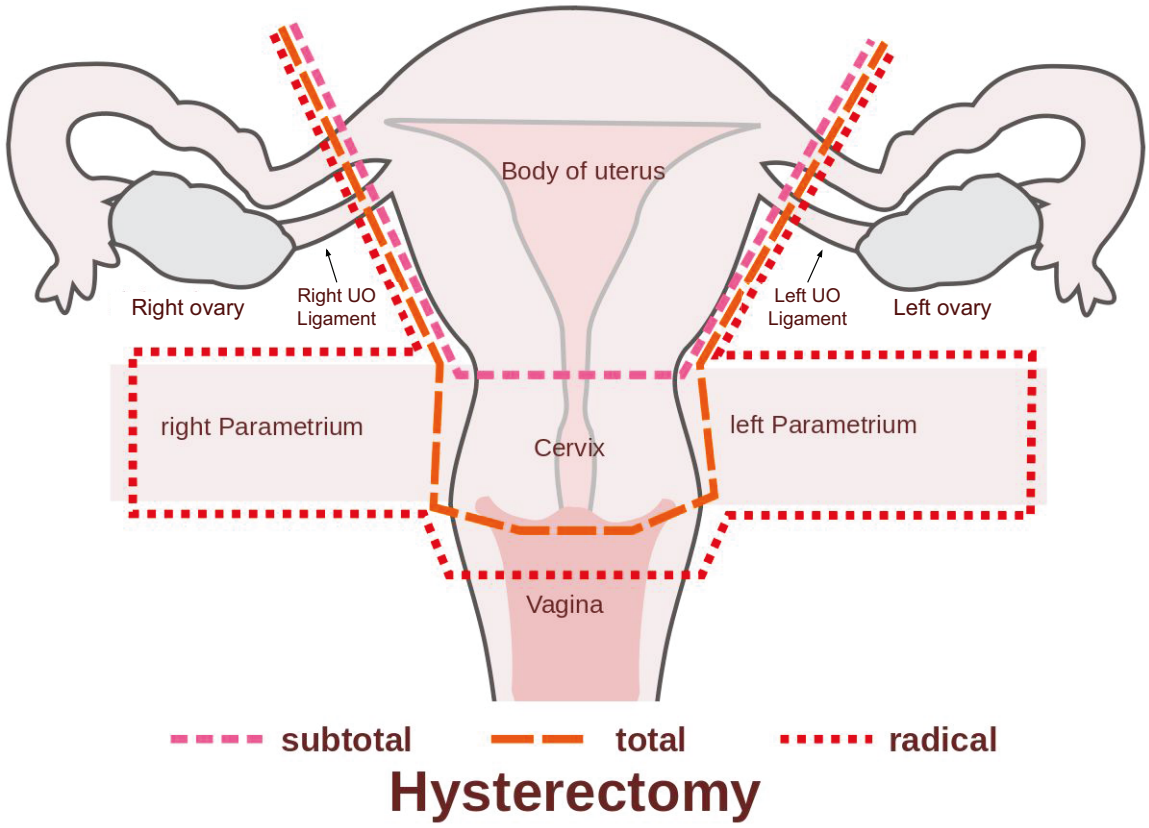


Figure A.5: Anatomical layout of the organs in Hysterectomy procedure

Task Description and Flow

The procedure of RALH is performed in women and involves surgical removal of uterus and cervix. Sometimes, this is accompanied by the removal of the ovaries (oophrectomy) and our data set contains procedures of that nature as well. Please refer to Figure A.5 for a pictorial description of anatomy around the uterus for better understanding of the task flow. Typically, RALH involves the setup phase of making incisions on the patient's belly for inserting the trocars, and is followed by the setup of the dVSS PSMs and instruments. Our data set skips this portion for privacy issues with a risk of capturing patient identifiers. Hysterectomy flow is relatively unstructured compared to cholecystectomy (gall bladder removal), which has a straight forward task flow. Generally, the surgeons isolated the uterus by dissecting the structures around it. First, we have the ligaments viz. infundibulopelvic (IP), round and utero-ovarian (UO). IP ligaments connect the ovaries to the pelvic wall on either side. Round ligaments connect the uterus to the pelvic walls on either side. UO ligaments connect the ovaries and uterus on either side. The surgeon may have to perform dissection to expose these ligaments. Following which the process of ligation occurs which is coagulation and sealing of the ligament before it can be cut. Note, that these ligaments can be dissected in any order the surgeon may prefer and at times a surgeon may perform half dissection of the ligament and proceed to some other step and come back to it later. After these primary ligaments, the dissection of the broad ligament begins. Anatomically, the above ligaments are part of the broad ligament,

which spreads around the uterus and connects it to the walls and floor of the pelvis. For future reference, we shall refer to the dissection of the tissue that is part of the broad ligament as the general step of 'Isolation of uterus'. Once the uterus has been isolated from the surrounding structures, the step of colopotomy is performed which involves the cutting of the cervical stump to open the vaginal canal. The uterus along with the cervix is then extracted through the vaginal canal. Suturing is performed to close and seal the vaginal cuff region. Our data set contains multiple types of suturing *viz.* running suture, interrupted suture, V-lock suture, and figure-of-eight suture. This is followed by clean up (haemostasis, suction and irrigation). Depending on the reason for hysterectomy, another additional step of lymph node dissection may be performed to run pathological tests. These typically occurred at the start or end of the procedure. The structure of the procedure flow is lost in the initial half of the procedure while isolating the uterus. Since the broad ligament spans most of the dissection region around the uterus, surgeons occasionally go back and forth between the different anatomical structures while isolating the uterus. This is quite evident in the task flow annotations described in the later sections.

Hardware

All the procedures were performed using the dVSS Si model. A variety of instruments were used depending on the surgical step and surgeon's preference. A monopolar curved scissors instrument was always used during dissection. This was

typically accompanied by a bipolar device like maryland bipolar forceps or fenestrated bipolar forceps. At least one of the variants of needle driver was used during the suturing step, sometimes two. A third arm instrument like prograsp forceps was used regularly as well. In addition to these EndoWrist™ instruments, traditional laparoscopic instruments like graspers, forceps, scissors, suction, retractors and organ retrievers were used by bed-side assistants to support the procedure. Some of the surgeons used the Firefly™ fluorescence imaging feature for lymph node dissection as well.

Study Participants

The study's main subjects were residents and 23 of them were recruited following an open call sent out to the residents from gynecology and obstetrics divisions. The surgical cases from six attending gynecologists were recorded as part of the study.

Study Protocol

This was a randomized trial design with the subjects acting as their own controls. The participants performed either two or four cases with one or two of them assigned as warm-up respectively. As part of the study participation, the residents were to familiarize themselves with the dVSS and dVSim. They were to perform and practice a minimum set of five VR training tasks viz. Energy Dissection I, Match Board I, Camera Targeting I, Suture Sponge I and Tubes using the dVSim Si model. The

dVSim was available 24/7 in a closet area outside the ORs for ease of access during warm-up. Based on the randomization, the residents were to perform a warm-up consisting of the above five exercises at least 15 minutes prior to the actual procedure. The participating attending surgeons had previously agreed to allow the residents to perform considerable portions of the procedure including some portion of dissection and some of the suturing. After the completion of the procedure, the residents were asked to fill out a questionnaire about what percentage of the different steps did they perform, a self evaluation using the OSATS and GEARS rating tools, and perspective about the warm-up and its effectiveness. The attending surgeons filled a questionnaire as well indicating similar fields as above and evaluating the resident's performance and indicating whether they felt the resident had performed a warm-up.

Data Collected

The da Vinci recorder tool from Section 2.1.4 was used for data collection. The da Vinci API kinematics and events along with the stereo endoscope video were captured. The data from the questionnaires was digitized and stored as well. Data from the dVSim were recorded as well, we will not be describing that here.

Data Annotated

Immediate skill ratings using OSATS and GEARS were obtained from attendings evaluating residents as well as residents doing a self-evaluation. A rigorous task flow

labeling was performed as described below.

Activity Labeling

After multiple rounds of consulting with our collaborating gynecologist, a vocabulary for task decomposition of the procedure was created with defined start and end points. Following this, one individual performed the labeling for the complete data set marking the start and end video timestamps for each constituent task step. Two individuals verified the labels and ambiguities were resolved by the three of them meeting together. A list of the constituent tasks and their definitions is in Table A.2. Task labels less than 45 seconds in length were consumed into the adjacent labels for a cleaner labeling. Some of the labeled procedure flows are shown in Figure A.6.

Size

A total of 33 procedures were recorded. Of these, 30 contain videos, and of these, 27 contain da Vinci API kinematics and events.

Access / Restrictions

The data set is under an active IRB study for preliminary analyses of the study's secondary aims. A public release would require an approval from the IRB committee. Additionally, the data set release will require approval from ISI for removal of possible proprietary content. We are planning to make the data set publicly available soon

Table A.2: WarmUp Hysterectomy Data Set Task Labels

Name	Description
Ligate {IP/Round/UO} {Left/Right}	Starts with grasping of the ligament structure by the bipolar device and ends when the structure is cut completely by the scissors.
Isolate Ovary Left/Right	Any dissection around the ovary with the goal to separate them from the pelvic wall
Isolate Uterus	Any dissection other than the Ligate task defined above with the goal to separate the uterus from the pelvic wall or floor including adhesion removal.
Colpotomy	Starts with the sharp dissection using scissors around the circumference of the cervical stump and ends with the complete cutting of the cervix.
Suture Vaginal Cuff	Starts with the first insertion of the needle on either of the sides of the vaginal opening and ends with the last suture knot placed before clean up.
Dissect Lymph Nodes	Any dissection around the pelvis walls with the goal of removing the lymph nodes. This overrides the above defined actions of Isolate Ovary and Isolate Uterus.
Dissect Auxiliary Tissue	Any dissection not covered by the above defined labels falls here.
Extract Anatomy	Process of removing the uterus plus cervix, ovaries, lymph nodes and other dissected tissue using the vaginal canal or laparoscopic ports.
Transition	Any segment of procedure without any constructive action (any of the above labels) being performed including idle time, endoscope clean up, suction and irrigation, instrument change.

Note: any labels less than 45 seconds in length were consumed into adjacent labels.

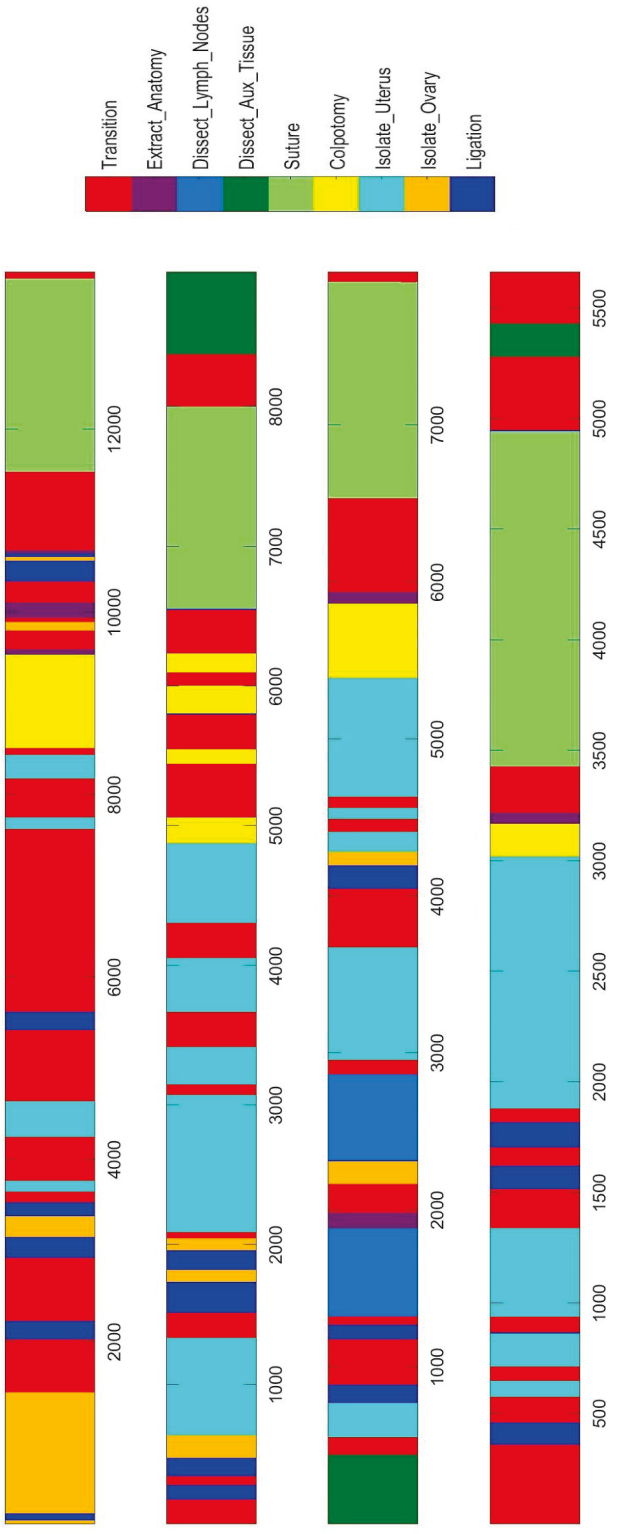


Figure A.6: Sample hysterectomy procedure flows with task labels.

after accomplishing the above goals.

Limitations / Pros

The data set is unique compared to other data sets available or previously mentioned in literature since it contains kinematics data from the dVSS. Previous OR studies have mainly focused on laparoscopic cholecystectomies for multiples reasons. One of them being it is a high volume procedure. But, more importantly it has a simple procedure flow for performing automated surgical activity recognition or skill assessment. Hysterectomy and prostatectomy are high volume procedures as well but have a complex workflow. While our data set contains multiple surgeons, they are all from the Johns Hopkins Medical Institution. Even then, there are striking differences in the approaches taken by different surgeons. However, including data of surgeons from other hospitals would be useful to claim broader applicability of methods validated using such a data set. Another characteristic of our data set is that every procedure is performed by two or three surgeons+residents making activity recognition and workflow analysis more challenging and realistic. Current skill ratings are missing in some cases and were not blinded since the attending surgeon was, of course, present in the room. This can be resolved by obtaining blinded reviews for portions of the surgery performed by the resident.

A.4 FESS Targeting Data Set

The functional endoscopic sinus surgery (FESS) data set was collected as part of a curricular training session for otolaryngology (head and neck) residents. We shall present some results using the data set in Chapter 4 and thus, we are including a short description here. The data set is like the ones described in[161,162] but not the same.

Goal

The goal of the data collection was to develop methods for computer-assisted path planning and skill assessment using instrument and endoscope motion and surgeon's eye tracking data.

Task Description

A partially dissected cadaver head was used in a training lab facility to perform FESS training tasks at the Johns Hopkins Hospital. Nine target locations were decided before the training lab began (listed in Table A.3). The target locations with respect to the cadaver being used are shown in Figure A.7.

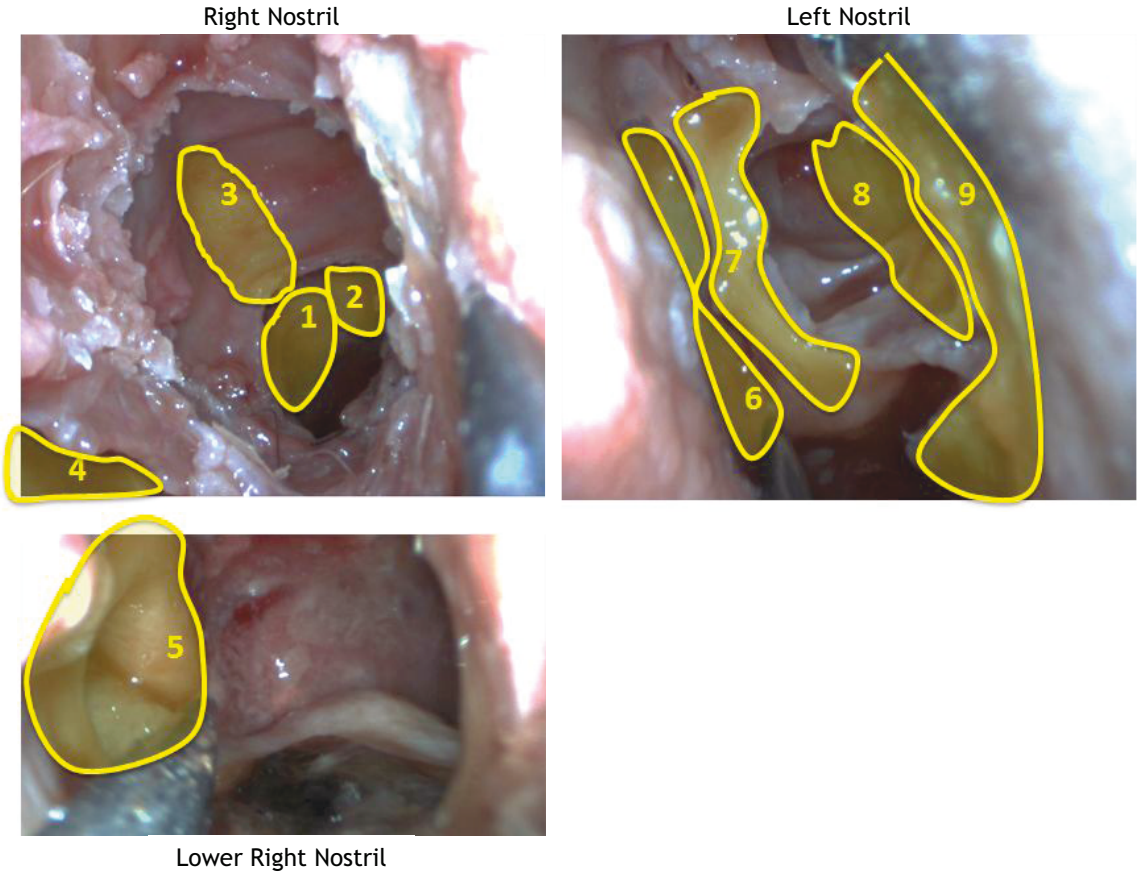


Figure A.7: Target locations for the FESS data set. Targets 1–4 are located in the right nostril, Target 5 is in the lower right nostril region, and Targets 6–9 are located in the left nostril. Refer to Table A.3 for names of the anatomical structure / regions.

Table A.3: FESS Data Set Target Anatomy

ID	Anatomical Name	Abbreviation	# of samples
1	right Carotid Artery	CA.r	47
2	right Sella Turcica	ST.r	46
3	right Optic Nerve	ON.r	48
4	right Maxillary Sinus	MS.r	49
5	right Eustachian Tube	ET.r	45
6	left Basal Lamella	BL.l	45
7	left Medial Wall of Bulla	MWB.l	21
8	left Lamina Papyracea	LP.l	49
9	left Uncinate Process	UP.l	21

Hardware

A nasal pointer instrument and a standard endoscope were provided to the participants to perform the task. Anatomical view from the endoscope was displayed on a 1200x800 resolution video monitor.

Task Flow

The participant entered the sinus cavity using the endoscope. They used the nasal pointer to locate and touch the specified target, and extracted the pointer and endoscope from the sinus cavity.

Study Participants

There were 20 participants in the study. Of these, 13 participants were novices (residents) and seven were experts (attending surgeons).

Study Protocol

This data was collected during a two day training lab for FESS. Each participant performed four test sessions - two on each day. A test session consisted of seven randomly chosen targets. The training lab instructor randomly selected the next target for each session impromptu. Thus, each participant was to perform a total of 14 targeting tasks. Every targeting task began and ended with the endoscope and nasal pointer instrument completely outside of the cadaver's sinus cavity.

Data Collected

Electromagnetic (EM) trackers were attached to the instrument and endoscope to record the 3D pose at 40Hz framerate along with a universal timestamp. An eye tracker device was used to record the participant's 2D gaze location on the video display at a 50Hz framerate. The video from the endoscope was recorded at 30Hz framerate and 1024x768 resolution as well.

Data Annotated

One individual segmented each test session into constituent targeting tasks by marking the start and end timestamps for each task. Three individuals evaluated task execution and provided binary and Likert-like ratings. Binary ratings were provided at the scale of each targeting task, while Likert-like rating on a scale of 1 to 5 was provided for the entire session consisting of a sequence of seven targeting tasks.

Size

A total of 51 sessions were recorded. Of these, 49 contain the motion, eye tracking and video data. A total of 371 targeting tasks across the nine targets were performed during these sessions. A distribution of the number of attempts per target is presented in Table A.3.

Access / Restrictions

We intend to make this data set public very soon. There are no IRB constraints and the data is striped off any identifiers.

Limitations / Pros

This data set is unique in terms of the different data modalities that were captured for the FESS training tasks *viz.* instrument and endoscope motion data, eye-gaze information, as well as the endoscope video. One limitation of the data set is lack

of rigor in randomization of ordering of the targets. A computer-generated ordering should be considered for a future study.

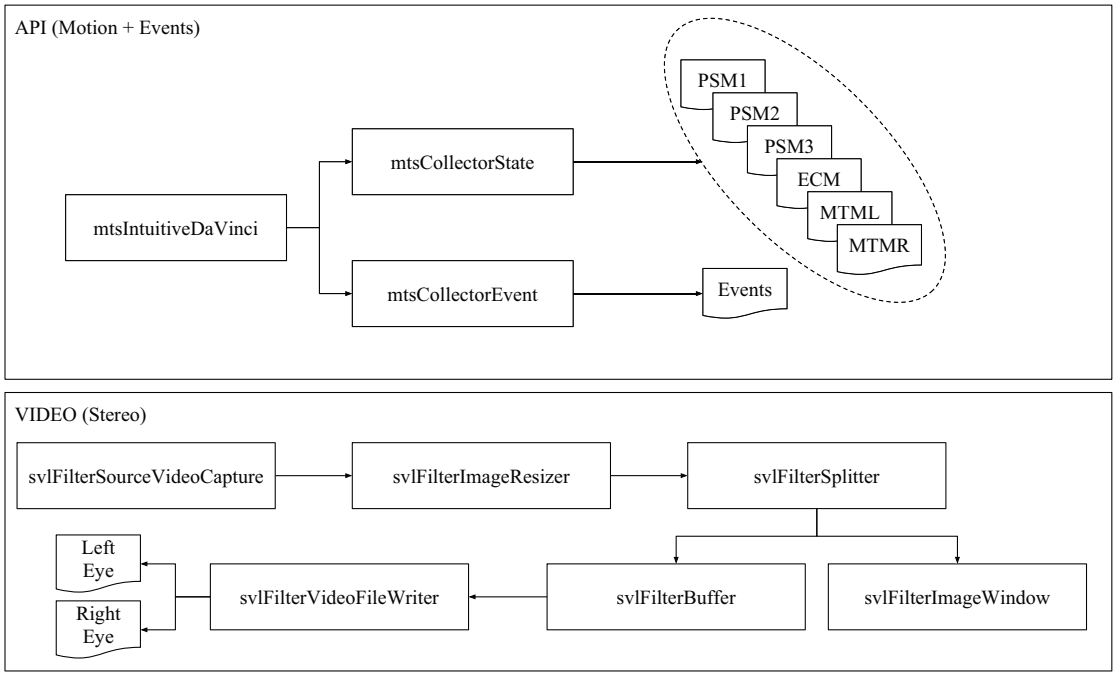


Figure B.1: da Vinci Recorder Tool: internal CISST and SAW components

Appendix B

da Vinci Recorder

B.1 Challenges in OR Data Collection

While the dVRecorder framework (Figure B.1) looks straightforward, there are multiple limitations and failure scenarios that can occur:

- ✖ The recording computer power cord is connected to one of the available receptacles in the OR. The OR staff can unplug the system if need arises to use the power socket for some other device. Thus, the person recording needs to understand to check for the incoming power.
- ✖ The UI access is strictly through the tablet device. The battery of the tablet should have enough charge to last the recording session in order to safely stop recording. A solution is to provide an alternative automatic start/stop on boot up of the recording computer.
- ✖ While it is rare, it is possible that the video output ports of the dVSS are being used for other external displays in the room. This situation is dependent on each new location where data is being captured.
- ✖ The storage racks in the vision cart of the dVSS might be already occupied by other devices like electrosurgery units and there may not be room for the recording computer. Reducing the form factor or making the dVRecorder an

integral part of the dVSS may solve this.

- ✖ The computer will have a limited amount of disk space, and so a routine data transfer using an external drive is required.
- ✖ It is possible that wireless network connectivity may be limited in the room or using wireless networks may be prohibited by the networking staff. In such a scenario, there will be no interface to access the UI.

Appendix C

Crowdsourcing

Definition

The authors Estellés-Arolas and González-Ladrón-de-Guevara in[116] built an integrated definition for crowdsourcing by surveying milestone works in the domain of crowdsourcing. The definition is quoted from the paper below:

> Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken.

C.1 Terminology

MTurk API

MTurk provides a friendly UI for conducting crowdsourcing tasks and managing them. However, our work involves surgical data which has not been widely explored in the crowdsourcing literature. Additionally, it involved complex survey layouts and hosting of data on internal storage servers. As a result of these, we chose to use the developer API (`http://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/Welcome.html`) provided by MTurk to create customized survey tools. We used the Python package `boto`[1] (`https://github.com/boto/boto`) that provides a wrapper `boto.mturk` around the MTurk API for easy use in a Python framework.

Requester

Any entity – an individual, an institution, a non-profit organization or company – that requires a crowd to provide a service to perform a specified task is termed as *requester*. As specified on MTurk's website, the name of the requester (as specified on their Amazon.com online account) is visible to the crowd performing the task. In addition to the general guidelines specified by MTurk, a group of academic researchers in collaboration with the crowd on MTurk have specified a guidelines document for

[1]Note: as of December 2016, version 3 of boto: `boto3` does not contain the `mturk` sub-module

academic research performed using MTurk.[217] We followed these guidelines while conducting all of our work on MTurk.

Worker

An individual who is a member of the crowd undertaking a task specified by the requester is a *worker*. The worker and requester together have some form of mutual benefit arising from crowdsourcing. On MTurk, while not primary, a definite benefit for the worker is reimbursement in the form of money paid by the requester. In other scenarios like crowdfunding the crowd brings money to support the requester's financial cause. Currently, a total of 500,000+ workers from across 190 countries are registered on MTurk, while majority of these are from U.S., followed by India. Please refer to this online tool[218] (`http://demographics.mturk-tracker.com/`) that publishes current demographics of MTurk workers including nationality, gender, primary income, household size for further details. Additionally, MTurk assigns a unique worker id that matches the regular expression: "`A[A-Z,0-9]{14,}`" - which means it a 14 or more character long alpha-numeric string starting with the letter 'A'.

Online Forums

In addition to the MTurk portal, workers have created forums and communities to interact, initiate, learn, and discuss the various requesters and tasks as well. These forums also have separate threads for requesters to be available for answering ques-

tions in case things need clarifications. MTurk also provides a short message tool to contact a requester. A noteworthy website is Turkopticon (TO) wherein workers can submit reviews about requesters whose tasks they have undertaken. This was created by Irani et al.[219] to enable workers to engage one another in mutual aid and evaluate their relation with the employers (requesters). As far as our experience goes, these reviews and ratings are important for faster acceptance of one's tasks by workers. Similarly, the workers maintain a blacklist too on most of these communities to make aware new workers about requesters who have played foul previously.

C.2 Human Intelligent Task (HIT) and Attributes

The basic premise of crowdsourcing is to utilize human intelligence in a collective manner to perform certain tasks which are thus, referred as *human intelligent tasks* (HITs). On MTurk, a requester posts a HIT to be performed by workers and this HIT appears on the worker portal as shown in Figure 2.18 along with some of its attributes (title, duration, etc.). MTurk assigns a unique identifier to each HIT which is typically a 30 character long alphanumeric string. HITs can be of a variety of types but belong to three broad categories – surveys, solving problems, and designing and creating content. A requester can specify a set of attributes for each HIT as follows. A screen capture of a sample HIT from one of our surveys is shown in Figure C.1.

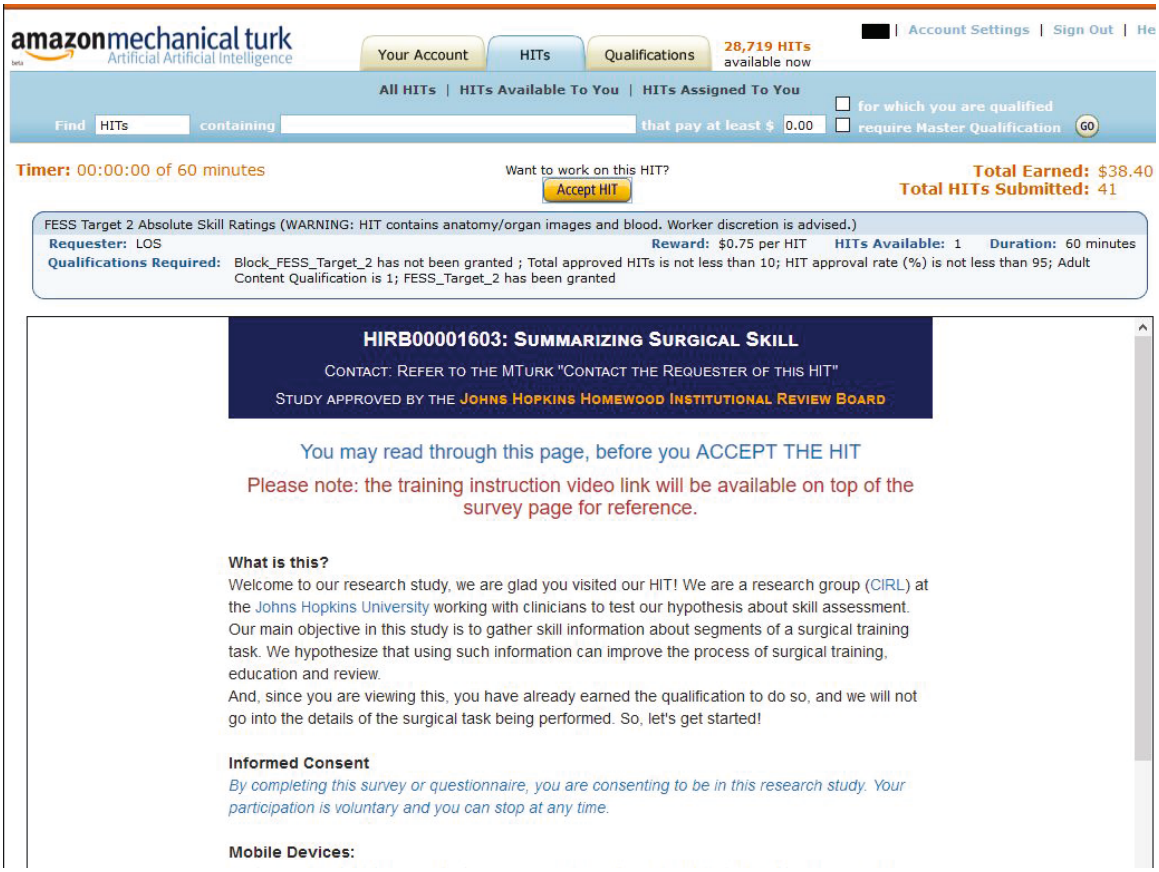


Figure C.1: A screen capture showing a sample HIT page with an external web page embedded in the MTurk web page.

Lifetime

This is the amount of time (since creation of the HIT) for which the HIT will appear on the MTurk portal. Although the maximum lifetime value possible is 365 days, in our experience HITs with longer lifetimes get pushed to later pages on the portal listing. Observing this, we adapted and started publishing HITs with a lifetime of 24 hours.

Duration

This is the amount of time that is allotted to a worker to finish the HIT. MTurk API allows HIT duration values in the range of 30 seconds to 365 days. In our case, we estimated the duration by performing the HITs within our research group and taking the average. However, this was a bad estimate since none of us are regular MTurk workers. On MTurk, a worker might be working on 25 HITs at a time, which means they will not be able to complete our HIT in the estimate that was based on a dedicated internal research group member. Thus, we allotted at least four-five times the estimated time as the HIT duration attribute and updated it according to initial completion times.

Question Form

While the MTurk API provides a set of question data structures for creating a web-based form that is hosted by MTurk, our studies include private surgical data as well as human subjects data. As per IRB guidelines and concerns to protect the privacy of the data set, we chose the alternative provided by the API to create an 'External Question' form. In this type, a requester can host the web-based survey or task on their privately managed web server, and MTurk will embed the external web page as an `iframe` onto the MTurk HIT web page. All of our work was using the external question type. The requester specifies the URL for the survey as a parameter to the question form in the API.

Assignment

Another term associated with MTurk is an *assignment*, which is basically acceptance of a HIT by a worker. We can think of it as a short term (micro) contract between requester and worker through the middle agent (MTurk). MTurk generates a unique identifier for each assignment that is typically 30-character long string and associated with the worker's ID and HIT's ID. A requester can specify the maximum number of assignments for each HIT they publish on the portal. MTurk makes sure that a worker can accept the assignment for a HIT only once. There are no restrictions from MTurk on acceptance of multiple different HITs by a worker other than

an upper limit of 25 HITs at any given time. Each assignment generated by MTurk can have the following outcomes:

I. Submission

This occurs when a worker accepts (before the HIT's *lifetime*) and completes the assignment within the allotted time (HIT *duration*). A submitted assignment has to be reviewed by the requester compulsorily within a specified time (an attribute of the HIT *auto approval delay* specified by the requester while creating the HIT). The requester has two choices: *Approval* and *Rejection*. The worker gets paid upon approval and vice versa.

II. Expiration

This occurs when a worker accepts the assignment but is not able to complete the assignment in the specified time (HIT duration). A countdown timer starts running on top of the HIT window indicating the amount of time remaining upon acceptance of the assignment (Figure C.1). Upon expiration of the timer, MTurk automatically redirects the worker to the portal page with a message indicating that the assignment has expired. An expired assignment becomes available to other workers.

III. Return

This occurs when a worker accepts the assignment but later decides that he/she doesn't want to finish it for whatever reason may be. MTurk provides a `Return HIT`

button on top of the accepted HIT assignment window for this case. A returned assignment becomes available to other workers.

At the end (of the HIT's lifetime) it is possible that there are some assignments remaining for a HIT. MTurk refunds the money back to the requester for the remaining assignments.

Reward

We mentioned that an assignment is a micro contract between the worker and requester. Thus, like any other contract, a pre-defined amount must be specified by the requester while creating a HIT which is referred to as *reward*. MTurk is the agent handling this contract and deducts the reward multiplied by the number of assignments requested for the HIT from the requester's pre-paid MTurk account. Upon the requester's approval of a submitted assignment, MTurk pays the pre-specified reward to the corresponding worker.

Amazon's Fees

In addition to this amount, MTurk charges a 20% of the reward or $0.01 (whichever is larger) as service fee for being the agent. An additional 20% fee is charged if the number of assignments requested per HIT is greater than or equal to 10. Thus, it is recommended and we set a maximum of nine assignments per HIT. In this case, we launched multiple HITs of the same type and made sure that the same worker could

not attempt the same HIT repeatedly. Another small caveat is that Amazon rounds up the fee charged to two decimal places such that $0.245 is rounded up to $0.25 and $0.244 is rounded down to $0.24.

Fair Wages

With contract, payment and duration of assignment, comes another important ethical issue of fairness in payment. MTurk workers have complained and discussed about low paying HITs and requesters across their online forums and communities. Almost all of them feature a top listed topic "**Which requesters to avoid?**". During the initial years of MTurk such communities were absent and some requesters did exploit workers by either paying them too low of rewards or rejecting completed work without giving rational reasons. As per the guidelines on most forums and the signed document by academic requesters and workers, a $6 per hour rate is currently accepted as standard.[217]

Bonus

While not an attribute of a HIT, MTurk allows requesters to pay bonuses to workers. One reason to grant a bonus would be to motivate the worker to accept and complete more assignments from the requester in future. A similar 20% fee (or $0.01 whichever is larger) is applied by MTurk to each bonus payment made by the requester.

In addition to these attributes, a HIT has a *title*, *description*, set of *keywords*, and *auto approval delay* (time after which an assignment is automatically accepted).

C.3 Qualifications

An important concept associated with crowdsourcing and with MTurk is that of *qualifications*. Since there are 500,000+ workers on MTurk and anyone else can join the pool by just signing up online, a requester can limit the set of workers who can work on their HITs. MTurk provides a pre-existing set of qualification types. We will explain the more relevant and commonly used ones below. Additionally, MTurk provides a special group of workers who have been assigned the `Master` qualification. As per MTurk, these are workers who have shown excellence across a wide range of HITs over a period of time. Of course, MTurk charges an additional fee of 5% (on top of the 20%) for using Master workers (please refer to `https://requester.mturk.com/pricing` for further details). A worker can browse through available qualifications on the portal and request qualifications.

Adult Content Qualification

MTurk assigns this qualification if the worker agrees that they are above 18 years of age and are willing to participate in HITs that may contain adult content. We used this qualification in all our HITs due to the nature of surgical videos and ethics

requirements specified in our IRB protocol for recruiting adults while conducting human subjects research.

Total HITs Approved

MTurk updates a count for each worker indicating the number of HITs they have completed and have been approved by the respective requesters. Such a qualification allows a requester to select more experience workers compared to someone who might have joined very recently. We used a threshold of 100 and above for all our work based on a previous work by Chen et al.[23]

HIT Approval Rate

MTurk updates a percentage value indicating the fraction of HITs that are approved out of the total that are submitted by a worker in the past. This value lies in the interval 0 to 100. This qualification is typically combined with the 'Total HITs Approved' qualification to filter workers who have some experience and that experience is good. We used a threshold of 95 and above for all our work based on Chen et al.[23]

Custom Qualification

In addition to the pre-existing qualifications, requesters can create their own qualifications to further filter the pool of workers for their HITs. A custom qualification

would typically involve a test. Workers that pass the test are granted the qualification. At times, a training/orientation section might be included by the requester to explain the content and concept related to the HITs. Since, watching and summarizing surgical videos are not usual HITs, all our work required custom qualifications as well. Training and orientation for our HITs required 3–5 minutes. However, MTurk does not allow hosting external web pages for qualifications.

We followed the approach outline in Figure C.2 to train and qualify workers to perform our HITs. We created a separate qualification HIT for our surveys. A worker would accept the HIT, watch a training video, read some instructions, and answer a test. Irrespective of their responses, we paid them a small reward. Workers who passed the test were granted a custom qualification along with a bonus payment for their effort.

Block Qualifications

While the MTurk API presents a method to block workers from attempting one's HITs, we were informed at the very start to avoid doing so. Apparently, Amazon starts blacklisting workers if they receive more blocks from requesters. Instead, it is recommended to create a blocking qualification and assign it to the workers who should not be attempting any more HITs posted by the requester. While creating such new HITs, just add a condition that the block qualification should not exist for the workers allowed to attempt it.

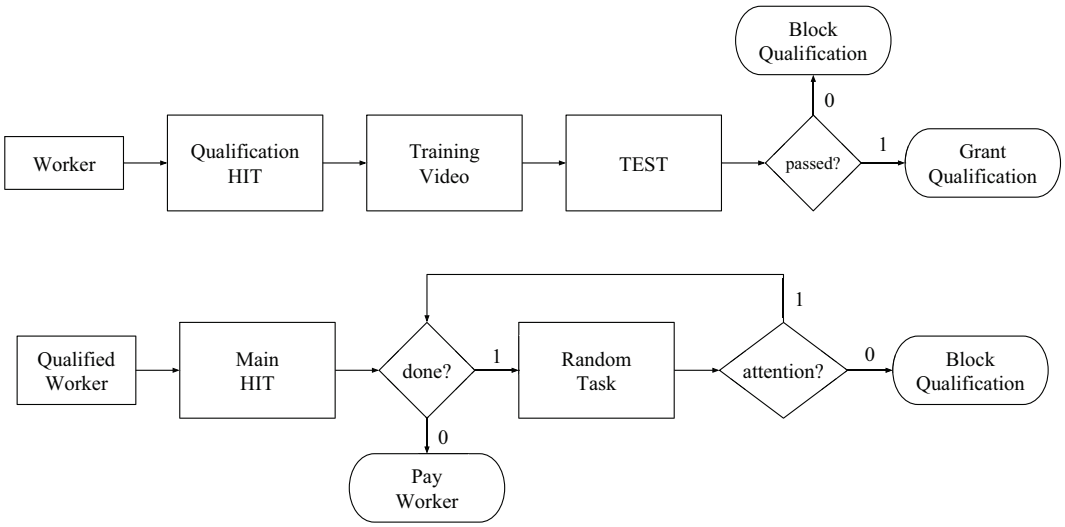


Figure C.2: Flow diagram of our crowdsourcing approach to train, orient, test and qualify workers for surgical HITs

C.4 Other Terminology

Experts

In addition to crowd workers performing the HITs specified by us, we asked a group of experts to perform the same HITs. This was important to obtain a set of reference annotations / responses to test the validity of the crowd's collective wisdom in the HITs published by us. Experts referred to attending surgeons with experience in performing the surgical tasks or procedures that were shown as part of the HITs. In this case as well, we collected responses from multiple experts on each task within our HITs to account for personal bias. We will refer to this process as *expertsourcing* and the set of responses as *expertsourced* responses.

Responses

All the HITs from our work contained surveys with multiple choice questions. The only open-ended questions were feedback and comments that we obtained from the workers. We will refer to these multiple-choice answers as *responses*, henceforth. The success of crowdsourcing is in the use of independent and diverse crowd to answer the HITs. All of our HITs collected responses from multiple workers (and experts), which means that in order to choose one collective response, we have to adopt a pooling strategy. We refer to process of combining multiple responses to obatin a single response on a question of the HIT as *pooling*. Depending on the type of measurement of the questions – binary (nominal), ordinal, interval and ratio – different pooling techniques may be employed.

Irrespective of whether we conduct the crowdsourcing studies and surveys on MTurk or using a university-wide recruitment we will use the above terms for concepts they refer to in later portions of this thesis.

Appendix D

Glossary of Terms

D.1 Abbreviations

Below we have listed all the commonly occurring acronyms and abbreviations used in this thesis for quick reference.

Abbreviation	Description
3D	Three Dimension(al)
ACGME	Accreditation Council for Graduate Medical Education
API	Application Programming Interface
AT	Annotation Task
CI	Confidence Interval
CIS	computer integrated surgery/surgical
CQ	Custom Qualification
DP	Deliberate Practice
dVSim	da Vinci® Skills Simulator™
dVSS	da Vinci® Surgical System
dVRecorder	da Vinci Data Recorder
ECM	Endoscope Camera Manipulator
ESU	Electro Surgery Unit
FESS	Functional Endoscopic Sinus Surgery
GEARS	Global Evaluative Assessment of Robotic Skills
GOALS	Global Operative Assessment of Laparoscopic Skills
GRS	Global Rating Scores
HIT	Human Intelligent Task
ISI	Intuitive Surgical Inc.
IRB	Institutional Review Board
JHU	Johns Hopkins University
MIS	Minimally Invasive Surgery
MTM	Master Tele Manipulator
MTurk	Amazon Mechanical Turk
NP	Needle Pass(ing)
OSATS	Objective Structured Assessment of Technical Skills
OR	operating room
PSC	Patient Side Cart (da Vinci)
PSM	Patient Side Manipulator (da Vinci)
RAMIS	Robot Assisted Minimally Invasive Surgery
RCM	Remote Center of Motion
RCT	Randomized Controlled Trial
RF	Random Forests
SSC	Surgeon Side Console (dVSS)
SG	SenseGraphics AB
SVM	Support Vector Machine
UI	User Interface
VR	virtual reality
VC	(Automated) Virtual Coach

Bibliography

[1] J. A. Martin, G. Regehr, R. Reznick, H. MacRae, J. Murnaghan, C. Hutchison, and M. Brown, "Objective structured assessment of technical skill (OSATS) for surgical residents," *Br J Surg*, vol. 84, no. 2, pp. 273–278, Feb. 1997.

[2] A. C. Goh, D. W. Goldfarb, J. C. Sander, B. J. Miles, and B. J. Dunkin, "Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills," *J Urol*, vol. 187, no. 1, pp. 247–252, Jan. 2012.

[3] I. Philibert, "New Requirements for Resident Duty Hours," *JAMA*, vol. 288, no. 9, p. 1112, Sep. 2002. [Online]. Available: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.288.9.1112

[4] J. H. Peters, G. M. Fried, L. L. Swanstrom, N. J. Soper, L. F. Sillin, B. Schirmer, K. Hoffman, and t. S. F. Committee, "Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery," *Surgery*, vol. 135, no. 1, pp.

21–27, Jan. 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0039606003001569

[5] B. J. Dunkin, "Fundamentals of Endoscopic Surgery," in *The SAGES Manual of Quality, Outcomes and Patient Safety*, D. S. T. FACS, MD, J. M. MPH, MD, and D. B. Jones, Eds. Springer US, 2012, pp. 473–484, dOI: 10.1007/978-1-4419-7901-8_47. [Online]. Available: http://link.springer.com/chapter/10.1007/978-1-4419-7901-8_47

[6] R. Smith, V. Patel, and R. Satava, "Fundamentals of robotic surgery: a course of basic robotic surgery skills based upon a 14-society consensus template of outcomes measures and curriculum development," *Int J Med Robotics Comput Assist Surg*, vol. 10, no. 3, pp. 379–384, Sep. 2014. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/rcs.1559/abstract

[7] D. Mutabdzic, M. Mylopoulos, M. L. Murnaghan, P. Patel, N. Zilbert, N. Seemann, G. Regehr, and C.-A. Moulton, "Coaching Surgeons: Is Culture Limiting Our Ability to Improve?" *Annals of Surgery*, vol. 262, no. 2, pp. 213–216, Aug. 2015.

[8] A. Gawande, "Personal Best," *The New Yorker*, Oct. 2011. [Online]. Available: http://www.newyorker.com/reporting/2011/10/03/111003fa_fact_gawande

[9] H. Min, D. R. Morales, D. Orgill, D. S. Smink, and S. Yule, "Systematic review of coaching toenhance surgeons' operative performance,"

Surgery, vol. 158, no. 5, pp. 1168–1191, Nov. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0039606015001695

[10] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal Convolutional Networks: A Unified Approach to Action Segmentation," *arXiv:1608.08242 [cs]*, Aug. 2016, arXiv: 1608.08242. [Online]. Available: http://arxiv.org/abs/1608.08242

[11] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos," *arXiv:1602.03012 [cs]*, Feb. 2016, arXiv: 1602.03012. [Online]. Available: http://arxiv.org/abs/1602.03012

[12] L. Chen, Q. Zhang, P. Zhang, and B. Li, "Instructive video retrieval for surgical skill coaching using attribute learning," in *2015 IEEE International Conference on Multimedia and Expo (ICME)*, Jun. 2015, pp. 1–6.

[13] Y. Gao, S. S. Vedula, G. I. Lee, M. R. Lee, S. Khudanpur, and G. D. Hager, "Query-by-example surgical activity detection," *Int J CARS*, pp. 1–10, Apr. 2016. [Online]. Available: http://link.springer.com/article/10.1007/s11548-016-1386-3

[14] A. Kapoor and R. H. Taylor, "A constrained optimization approach to virtual fixtures for multi-handed tasks," in *IEEE International Conference on Robotics and Automation, 2008. ICRA 2008*, May 2008, pp. 3401–3406.

[15] Z. Chen, A. Malpani, P. Chalasani, A. Deguet, S. S. Vedula, P. Kazhanzides, and R. H. Taylor, "Virtual fixture assistance for needle passing and knot tying," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2016, pp. 2343–2350.

[16] N. Padoy and G. D. Hager, "Spatio-Temporal Registration of Multiple Trajectories," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2011*, ser. Lecture Notes in Computer Science, G. Fichtinger, A. Martel, and T. Peters, Eds. Springer Berlin Heidelberg, Jan. 2011, no. 6891, pp. 145–152. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-23623-5_19

[17] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popovi, and F. Players, "Predicting protein structures with a multiplayer online game," *Nature*, vol. 466, no. 7307, pp. 756–760, Aug. 2010. [Online]. Available: http://www.nature.com/nature/journal/v466/n7307/full/nature09304.html

[18] R. Kffner, N. Zach, R. Norel, J. Hawe, D. Schoenfeld, L. Wang, G. Li, L. Fang, L. Mackey, O. Hardiman, M. Cudkowicz, A. Sherman, G. Ertaylan, M. Grosse-Wentrup, T. Hothorn, J. van Ligtenberg, J. H. Macke, T. Meyer, B. Schlkopf, L. Tran, R. Vaughan, G. Stolovitzky, and M. L. Leitner, "Crowdsourced analysis of clinical trial data to predict amyotrophic lateral

sclerosis progression," *Nat Biotech*, vol. 33, no. 1, pp. 51–57, Jan. 2015. [Online]. Available: http://www.nature.com/nbt/journal/v33/n1/full/nbt.3051.html

[19] K. Abdallah, C. Hugh-Jones, T. Norman, S. Friend, and G. Stolovitzky, "The Prostate Cancer DREAM Challenge: A Community-Wide Effort to Use Open Clinical Trial Data for the Quantitative Prediction of Outcomes in Metastatic Prostate Cancer," *The Oncologist*, vol. 20, no. 5, pp. 459–460, May 2015. [Online]. Available: http://theoncologist.alphamedpress.org/content/20/5/459

[20] F. J. Candido dos Reis, S. Lynn, H. R. Ali, D. Eccles, A. Hanby, E. Provenzano, C. Caldas, W. J. Howat, L.-A. McDuffus, B. Liu, F. Daley, P. Coulson, R. J. Vyas, L. M. Harris, J. M. Owens, A. F. M. Carton, J. P. McQuillan, A. M. Paterson, Z. Hirji, S. K. Christie, A. R. Holmes, M. K. Schmidt, M. Garcia-Closas, D. F. Easton, M. K. Bolla, Q. Wang, J. Benitez, R. L. Milne, A. Mannermaa, F. Couch, P. Devilee, R. A. E. M. Tollenaar, C. Seynaeve, A. Cox, S. S. Cross, F. M. Blows, J. Sanders, R. de Groot, J. Figueroa, M. Sherman, M. Hooning, H. Brenner, B. Holleczek, C. Stegmaier, C. Lintott, and P. D. P. Pharoah, "Crowdsourcing the General Public for Large Scale Molecular Pathology Studies in Cancer," *EBioMedicine*, vol. 2, no. 7, pp. 681–689, Jul. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352396415300165

[21] S. Mavandadi, S. Dimitrov, S. Feng, F. Yu, U. Sikora, O. Yaglidere,

S. Padmanabhan, K. Nielsen, and A. Ozcan, "Distributed Medical Image Analysis and Diagnosis through Crowd-Sourced Games: A Malaria Case Study," *PLoS ONE*, vol. 7, no. 5, p. e37245, May 2012. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0037245

[22] L. I. Lesser, L. Wu, T. B. Matthiessen, and H. S. Luft, "Evaluating the healthiness of chain-restaurant menu items using crowdsourcing: a new method," *Public Health Nutrition*, pp. 1–7, Jan. 2016. [Online]. Available: https://www.cambridge.org/core/journals/public-health-nutrition/article/evaluating-the-healthiness-of-chain-restaurant-menu-items-using-crowdsourcing-a-new-meth 70E8B393B3C43B4DFFCAF608EA06B91C

[23] C. Chen, L. White, T. Kowalewski, R. Aggarwal, C. Lintott, B. Comstock, K. Kuksenok, C. Aragon, D. Holst, and T. Lendvay, "Crowd-Sourced Assessment of Technical Skills: a novel method to evaluate surgical performance," *J Surg Res*, vol. 187, no. 1, pp. 65–71, Mar. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022480413008998

[24] K. R. Ghani, D. C. Miller, S. Linsell, A. Brachulis, B. Lane, R. Sarle, D. Dalela, M. Menon, B. Comstock, T. S. Lendvay, J. Montie, and J. O. Peabody, "Measuring to Improve: Peer and Crowd-sourced Assessments of Technical Skill with Robot-assisted Radical Prostatectomy," *European*

Urology, vol. 69, no. 4, pp. 547–550, Apr. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0302283815011896

[25] R. W. Tse, E. Oh, J. S. Gruss, R. A. Hopper, and C. B. Birgfeld, "Crowdsourcing as a Novel Method to Evaluate Aesthetic Outcomes of Treatment for Unilateral Cleft Lip:," *Plastic and Reconstructive Surgery*, vol. 138, no. 4, pp. 864–874, Oct. 2016. [Online]. Available: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00006534-201610000-00026

[26] J. H. Park, S. Mirhosseini, S. Nadeem, J. Marino, A. Kaufman, K. Baker, and M. Barish, "Crowdsourcing for Identification of Polyp-Free Segments in Virtual Colonoscopy Videos," *arXiv:1606.06702 [cs]*, Jun. 2016, arXiv: 1606.06702. [Online]. Available: http://arxiv.org/abs/1606.06702

[27] D. Mitry, K. Zutis, B. Dhillon, T. Peto, S. Hayat, K.-T. Khaw, J. E. Morgan, W. Moncur, E. Trucco, and P. J. Foster, "The Accuracy and Reliability of Crowdsource Annotations of Digital Retinal Images," *Transl Vis Sci Technol*, vol. 5, no. 5, Sep. 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5032847/

[28] L. Maier-Hein, S. Mersmann, D. Kondermann, S. Bodenstedt, A. Sanchez, C. Stock, H. G. Kenngott, M. Eisenmann, and S. Speidel, "Can Masses of Non-Experts Train Highly Accurate Image Classifiers?" in

Medical Image Computing and Computer-Assisted Intervention MICCAI 2014, ser. Lecture Notes in Computer Science, P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, Eds. Springer International Publishing, Jan. 2014, no. 8674, pp. 438–445. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-10470-6_55

[29] R. M. Satava, "Virtual reality surgical simulator," *Surg Endosc*, vol. 7, no. 3, pp. 203–205, May 1993. [Online]. Available: http://link.springer.com/article/10.1007/BF00594110

[30] N. E. Seymour, A. G. Gallagher, S. A. Roman, M. K. O'Brien, V. K. Bansal, D. K. Andersen, and R. M. Satava, "Virtual Reality Training Improves Operating Room Performance," *Ann Surg*, vol. 236, no. 4, pp. 458–464, Oct. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1422600/

[31] S. C. Hultman, "The (R)Evolution of Surgical Education in the Twenty-first Century: Replacing the Traditional Paradigm of "See One, Do One, Teach One" With Competency-Based, Lifelong Learning - C. Scott Hultman, MD, MBA, FACS - QMP Plastic Surgery Pulse," 2009. [Online]. Available: http://www.plasticsurgerypulsenews.com/4/article_dtl.php?QnCategoryID=39&QnArticleID=97&QnCurPage=1

[32] K. E. Fletcher, D. A. Reed, and V. M. Arora, "Systematic Review of

the Literature: Resident Duty Hours and Related Topics," Sep. 2009. [Online]. Available: https://www.acgme.org/Portals/0/PDFs/Resident_Duty_Hours_and_Related_Topics[1].pdf

[33] D. A. DaRosa and C. M. Pugh, "Error training: Missing link in surgical education," *Surgery*, vol. 151, no. 2, pp. 139–145, Feb. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0039606011004740

[34] A. K. Gardner, K. Abdelfattah, J. Wiersch, R. A. Ahmed, and R. E. Willis, "Embracing Errors in Simulation-Based Training: The Effect of Error Training on Retention and Transfer of Central Venous Catheter Skills," *Journal of Surgical Education*, vol. 72, no. 6, pp. e158–e162, Nov. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S193172041500207X

[35] J. H. Barsuk, W. C. McGaghie, E. R. Cohen, J. S. Balachandran, and D. B. Wayne, "Use of simulation-based mastery learning to improve the quality of central venous catheter placement in a medical intensive care unit," *J. Hosp. Med.*, vol. 4, no. 7, pp. 397–403, Sep. 2009. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/jhm.468/abstract

[36] H. Khouli, K. Jahnes, J. Shapiro, K. Rose, J. Mathew, A. Gohil, Q. Han, A. Sotelo, J. Jones, A. Aqeel, E. Eden, and E. Fried, "Performance of Medical Residents in Sterile Techniques During Central Vein Catheterization: Randomized Trial of Efficacy of Simulation-Based

Training," *Chest*, vol. 139, no. 1, pp. 80–87, Jan. 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0012369211600172

[37] B. Zendejas, D. A. Cook, J. Bingener, M. Huebner, W. F. Dunn, M. G. Sarr, and D. R. Farley, "Simulation-Based Mastery Learning Improves Patient Outcomes in Laparoscopic Inguinal Hernia Repair: A Randomized Controlled Trial," *Annals of Surgery*, vol. 254, no. 3, pp. 502–511, Sep. 2011. [Online]. Available: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00000658-201109000-00013

[38] T. Draycott, T. Sibanda, L. Owen, V. Akande, C. Winter, S. Reading, and A. Whitelaw, "Does training in obstetric emergencies improve neonatal outcome?" *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 113, no. 2, pp. 177–182, Feb. 2006. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1471-0528.2006.00800.x/abstract

[39] J. S. Ilgen, J. Sherbino, and D. A. Cook, "Technology-enhanced Simulation in Emergency Medicine: A Systematic Review and Meta-Analysis," *Acad Emerg Med*, vol. 20, no. 2, pp. 117–127, Feb. 2013. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/acem.12076/abstract

[40] A. K. Gardner, D. Nepomnayshy, C. Reickert, D. W. Gee, R. Brydges, J. R. Korndorffer Jr., D. J. Scott, and A. K. Sachdeva, "The value proposition of

simulation," *Surgery*, vol. 160, no. 3, pp. 546–551, Sep. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0039606016300290

[41] B. Zendejas, A. T. Wang, R. Brydges, S. J. Hamstra, and D. A. Cook, "Cost: The missing outcome in simulation-based medical education research: A systematic review," *Surgery*, vol. 153, no. 2, pp. 160–176, Feb. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0039606012003182

[42] L. Chang, J. Petros, D. T. Hess, C. Rotondi, and T. J. Babineau, "Integrating simulation into a surgical residency program," *Surg Endosc*, vol. 21, no. 3, pp. 418–421, Dec. 2006. [Online]. Available: http://link.springer.com/article/10.1007/s00464-006-9051-5

[43] K. W. van Dongen, W. A. van der Wal, I. H. M. B. Rinkes, M. P. Schijven, and I. A. M. J. Broeders, "Virtual reality training for endoscopic surgery: voluntary or obligatory?" *Surg Endosc*, vol. 22, no. 3, pp. 664–667, Mar. 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2262145/

[44] M. A. C. Zapf and M. B. Ujiki, "Surgical Resident Evaluations of Portable Laparoscopic Box Trainers Incorporated Into a Simulation-Based Minimally Invasive Surgery Curriculum," *SURG INNOV*, vol. 22, no. 1, pp. 83–87, Feb. 2015. [Online]. Available: http://sri.sagepub.com/content/22/1/83

[45] A. J. Becker, "It's Not What They Do, It's How They Do It: Athlete

Experiences of Great Coaching," *International journal of Sports Science & Coaching*, vol. 4, no. 1, pp. 93–119, Mar. 2009. [Online]. Available: http://spo.sagepub.com/content/4/1/93

[46] P. Crochet, R. Aggarwal, S. S. Dubb, P. Ziprin, N. Rajaretnam, T. Grantcharov, K. A. Ericsson, and A. Darzi, "Deliberate practice on a virtual reality laparoscopic simulator enhances the quality of surgical technical Skills:," *Ann Surg*, vol. 253, no. 6, pp. 1216–1222, Jun. 2011. [Online]. Available: http://ovidsp.tx.ovid.com/sp-3.13.1a/ovidweb.cgi?QS2=434f4e1a73d37e8ccc93173c62ce7b0d5dea8a2ba362037dd496991477420e4c1fb7a874812d7992e2

[47] E. M. Bonrath, N. J. Dedy, L. E. Gordon, and T. P. Grantcharov, "Comprehensive Surgical Coaching Enhances Surgical Skill in the Operating Room: A Randomized Controlled Trial," *Annals of Surgery*, vol. 262, no. 2, pp. 205–212, Aug. 2015. [Online]. Available: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00000658-201508000-00001

[48] C. C. Greenberg, H. N. Ghousseini, S. R. Pavuluri Quamme, H. L. Beasley, and D. A. Wiegmann, "Surgical Coaching for Individual Performance Improvement," *Ann Surg*, vol. 261, no. 1, pp. 32–34, Jan. 2015.

[49] D. A. Rogers, G. Regehr, and J. MacDonald, "A role for error training in surgical technical skill instruction and evaluation," *The American Journal*

of Surgery, vol. 183, no. 3, pp. 242–245, Mar. 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002961002007985

[50] A. M. Jarc, S. H. Shah, T. Adebar, E. Hwang, M. Aron, I. S. Gill, and A. J. Hung, "Beyond 2d telestration: an evaluation of novel proctoring tools for robot-assisted minimally invasive surgery," *J Robotic Surg*, vol. 10, no. 2, pp. 103–109, Feb. 2016. [Online]. Available: http://link.springer.com/article/10.1007/s11701-016-0564-1

[51] A. M. Jarc, A. A. Stanley, T. Clifford, I. S. Gill, and A. J. Hung, "Proctors exploit three-dimensional ghost tools during clinical-like training scenarios: a preliminary study," *World Journal of Urology*, Sep. 2016. [Online]. Available: http://link.springer.com/10.1007/s00345-016-1944-x

[52] M. Vassiliou, L. Feldman, C. Andrew, S. Bergman, K. Leffondr, D. Stanbridge, and G. Fried, "A global assessment tool for evaluation of intraoperative laparoscopic skills," *Am J Surg*, vol. 190, no. 1, pp. 107–113, Jul. 2005.

[53] V. Datta, A. Chang, S. Mackay, and A. Darzi, "The relationship between motion analysis and surgical technical assessments," *Am J Surg*, vol. 184, no. 1, pp. 70–73, Jul. 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002961002008917

[54] R. Kumar, A. Jog, B. Vagvolgyi, H. Nguyen, G. Hager, C. C. G. Chen, and D. Yuh, "Objective measures for longitudinal assessment of robotic surgery

training," *J Thorac Cardiovasc Surg*, vol. 143, no. 3, pp. 528–534, Mar. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022522311012748

[55] T. M. Kowalewski, L. W. White, T. S. Lendvay, I. S. Jiang, R. Sweet, A. Wright, B. Hannaford, and M. N. Sinanan, "Beyond task time: automated measurement augments fundamentals of laparoscopic skills methodology," *Journal of Surgical Research*, vol. 192, no. 2, pp. 329–338, Dec. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022480414005411

[56] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal, "Sparse hidden Markov models for surgical gesture classification and skill evaluation," in *Information Processing in Computer-Assisted Interventions*, P. Abolmaesumi, L. Joskowicz, N. Navab, and P. Jannin, Eds. Springer Berlin Heidelberg, Jan. 2012, pp. 167–177. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-30618-1_17

[57] J. Rosen, B. Hannaford, C. Richards, and M. Sinanan, "Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills," *IEEE Trans Biomed Eng*, vol. 48, no. 5, pp. 579–591, May 2001.

[58] C. E. Reiley and G. D. Hager, "Decomposition of Robotic Surgical Tasks: An

Analysis of Subtasks and Their Correlation to Skill," *Medical Image Computing and Computer-Assisted Intervention M2CAI Workshop*, 2009.

[59] A. Jog, B. Itkowitz, M. Liu, S. DiMaio, G. Hager, M. Curet, and R. Kumar, "Towards integrating task information in skills assessment for dexterous tasks in surgery and simulation," in *2011 IEEE International Conference on Robotics and Automation (ICRA)*, May 2011, pp. 5273–5278.

[60] N. Ahmidi, Y. Gao, B. Bejar, S. S. Vedula, S. Khudanpur, R. Vidal, and G. D. Hager, "String Motif-Based Description of Tool Motion for Detecting Skill and Gestures in Robotic Surgery," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2013*, ser. Lecture Notes in Computer Science, K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, Eds. Springer Berlin Heidelberg, Jan. 2013, no. 8149, pp. 26–33. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-40811-3_4

[61] N. Ahmidi, "Activity Decomposition and Skill Assessment for Dextrous Motions in Robotic and Minimally-Invasive Surgery," Ph.D., The Johns Hopkins University, Baltimore, Maryland, 2015.

[62] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, M. A. Clements, and I. Essa, "Automated Assessment of Surgical Skills Using Frequency Analysis," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M.

Wells, and A. F. Frangi, Eds. Springer International Publishing, Oct. 2015, no. 9349, pp. 430–438, dOI: 10.1007/978-3-319-24553-9_53. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-24553-9_53

[63] Y. Sharma, T. Plotz, N. Hammerld, S. Mellor, R. McNaney, P. Olivier, S. Deshmukh, A. McCaskie, and I. Essa, "Automated surgical OSATS prediction from videos," Apr. 2014, pp. 461–464.

[64] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, T. Ploetz, M. A. Clements, and I. Essa, "Automated video-based assessment of surgical skills for training and evaluation in medical schools," *Int J CARS*, vol. 11, no. 9, pp. 1623–1636, Aug. 2016. [Online]. Available: http://link.springer.com.proxy1.library.jhu.edu/article/10.1007/s11548-016-1468-2

[65] J. Brown, C. O'Brien, S. Leung, K. Dumon, D. Lee, and K. Kuchenbecker, "Using Contact Forces and Robot Arm Accelerations to Automatically Rate Surgeon Skill at Peg Transfer," *IEEE Transactions on Biomedical Engineering*, vol. PP, no. 99, pp. 1–1, 2016.

[66] T. P. Grantcharov, S. Schulze, and V. B. Kristiansen, "The impact of objective assessment and constructive feedback on improvement of laparoscopic performance in the operating room," *Surg Endosc*, vol. 21, no. 12, pp. 2240–2243, Apr. 2007. [Online]. Available: http://link.springer.com/article/10.1007/s00464-007-9356-z

[67] G. G. Hamad, M. T. Brown, and J. A. Clavijo-Alvarez, "Postoperative video debriefing reduces technical errors in laparoscopic surgery," *The American Journal of Surgery*, vol. 194, no. 1, pp. 110–114, Jul. 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002961007002553

[68] M. C. Porte, G. Xeroulis, R. K. Reznick, and A. Dubrowski, "Verbal feedback from an expert is more effective than self-accessed feedback about motion efficiency in learning new surgical skills," *Am J Surg*, vol. 193, no. 1, pp. 105–110, Jan. 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002961006006398

[69] I. Kruglikova, T. P. Grantcharov, A. M. Drewes, and P. Funch-Jensen, "The impact of constructive feedback on training in gastrointestinal endoscopy using high-fidelity Virtual-Reality simulation: a randomised controlled trial," *Gut*, vol. 59, no. 2, pp. 181–185, Feb. 2010.

[70] E. Boyle, M. Al-Akash, A. G. Gallagher, O. Traynor, A. D. K. Hill, and P. C. Neary, "Optimising surgical training: use of feedback to reduce errors during a simulated surgical procedure," *Postgrad Med J*, p. pgmj.2010.109363, Jun. 2011. [Online]. Available: http://pmj.bmj.com/content/early/2011/06/03/pgmj.2010.109363

[71] J. Strandbygaard, F. Bjerrum, M. Maagaard, P. Winkel, C. R. Larsen, C. Ringsted, C. Gluud, T. Grantcharov, B. Ottesen, and J. L. Sorensen,

"Instructor Feedback Versus No Instructor Feedback on Performance in a Laparoscopic Virtual Reality Simulator: A Randomized Trial," *Annals of Surgery*, vol. 257, no. 5, pp. 839–844, May 2013. [Online]. Available: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00000658-201305000-00009

[72] L. Ahlborg, M. Weurlander, L. Hedman, H. Nisell, P. G. Lindqvist, L. Fellnder-Tsai, and L. Enochsson, "Individualized feedback during simulated laparoscopic training: a mixed methods study," *Int J Med Educ*, vol. 6, pp. 93–100, Jul. 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4537795/

[73] C. J. Vaughn, E. Kim, P. O'Sullivan, E. Huang, M. Y. C. Lin, S. Wyles, B. J. A. Palmer, J. L. Pierce, and H. Chern, "Peer video review and feedback improve performance in basic surgical skills," *The American Journal of Surgery*, vol. 211, no. 2, pp. 355–360, Feb. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S000296101500608X

[74] C. E. Reiley, T. Akinbiyi, D. Burschka, D. C. Chang, A. M. Okamura, and D. D. Yuh, "Effects of visual force feedback on robot-assisted surgical task performance," *The Journal of Thoracic and Cardiovascular Surgery*, vol. 135, no. 1, pp. 196–202, Jan. 2008. [Online]. Available: http://www.jtcvsonline.org/article/S002252230701536X/abstract

[75] D. A. Hashimoto, P. Sirimanna, E. D. Gomez, L. Beyer-Berjot, K. A.

Ericsson, N. N. Williams, A. Darzi, and R. Aggarwal, "Deliberate practice enhances quality of laparoscopic surgical performance in a randomized controlled trial: from arrested development to expert performance," *Surg Endosc*, vol. 29, no. 11, pp. 3154–3162, Dec. 2014. [Online]. Available: http://link.springer.com/article/10.1007/s00464-014-4042-4

[76] D. A. Davis, P. E. Mazmanian, M. Fordis, R. V. Harrison, K. E. Thorpe, and L. Perrier, "Accuracy of Physician Self-assessment Compared With Observed Measures of Competence: A Systematic Review," *JAMA*, vol. 296, no. 9, pp. 1094–1102, Sep. 2006. [Online]. Available: http://jamanetwork.com/journals/jama/fullarticle/203258

[77] R. H. Bell Jr., "Why Johnny cannot operate," *Surg*, vol. 146, no. 4, pp. 533–542, Oct. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0039606009004620

[78] P. Singh, R. Aggarwal, M. Tahir, P. H. Pucher, and A. Darzi, "A Randomized Controlled Study to Evaluate the Role of Video-based Coaching in Training Laparoscopic Skills:," *Ann Surg*, vol. 261, no. 5, pp. 862–869, May 2015.

[79] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93s1," 1993.

[80] J. Godfrey and E. Holliman, "Switchboard-1 Release 2 LDC97s62," 1993.

[81] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a Large Annotated Corpus of English: The Penn Treebank," *Comput. Linguist.*, vol. 19, no. 2, pp. 313–330, Jun. 1993. [Online]. Available: http://dl.acm.org/citation.cfm?id=972470.972475

[82] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "RCV1: A New Benchmark Collection for Text Categorization Research," *J. Mach. Learn. Res.*, vol. 5, pp. 361–397, Dec. 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1005332.1005345

[83] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *Int J Comput Vis*, vol. 88, no. 2, pp. 303–338, Jun. 2010. [Online]. Available: http://link.springer.com/article/10.1007/s11263-009-0275-4

[84] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Int J Comput Vis*, vol. 115, no. 3, pp. 211–252, Dec. 2015. [Online]. Available: http://link.springer.com/article/10.1007/s11263-015-0816-y

[85] D. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. Curtiss, S. Della Penna, D. Feinberg, M. Glasser, N. Harel, A. Heath, L. Larson-Prior, D. Marcus, G. Michalareas,

S. Moeller, R. Oostenveld, S. Petersen, F. Prior, B. Schlaggar, S. Smith, A. Snyder, J. Xu, and E. Yacoub, "The Human Connectome Project: A data acquisition perspective," *Neuroimage*, vol. 62, no. 4, pp. 2222–2231, Oct. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3606888/

[86] A. M. Wood, I. R. White, and S. G. Thompson, "Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals," *Clin Trials*, vol. 1, no. 4, pp. 368–376, Aug. 2004. [Online]. Available: http://ctj.sagepub.com/content/1/4/368

[87] R. M. Satava, "The Operating Room of the Future: Observations and Commentary," *SURG INNOV*, vol. 10, no. 3, pp. 99–105, Sep. 2003. [Online]. Available: http://sri.sagepub.com/content/10/3/99

[88] R. Bharathan, R. Aggarwal, and A. Darzi, "Operating room of the future," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 27, no. 3, pp. 311–322, Jun. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1521693412001794

[89] L. Maier-Hein, S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, M. Hashizume, D. Katic, H. Kenngott, M. Kranzfelder, A. Malpani, K. Mrz, T. Neumuth, N. Padoy, C. Pugh, N. Schoch, D. Stoyanov, R. Taylor, M. Wagner, G. D. Hager, and P. Jannin, "Surgical Data Science: Enabling Next-Generation

Surgery," *arXiv:1701.06482 [cs]*, Jan. 2017, arXiv: 1701.06482. [Online]. Available: http://arxiv.org/abs/1701.06482

[90] S. S. Vedula, M. Ishii, and G. D. Hager, "Perspectives on Surgical Data Science," *arXiv:1610.04276 [cs]*, Oct. 2016, arXiv: 1610.04276. [Online]. Available: http://arxiv.org/abs/1610.04276

[91] J. E. Wickham, "The new surgery." *British Medical Journal (Clinical research ed.)*, vol. 295, no. 6613, pp. 1581–1582, Dec. 1987. [Online]. Available: http://www.bmj.com/cgi/doi/10.1136/bmj.295.6613.1581

[92] P. Cain, P. Kazanzides, J. Zuhars, B. Mittelstadt, and H. Paul, "Safety considerations in a surgical robot," *Biomed Sci Instrum*, vol. 29, pp. 291–294, 1993.

[93] J. M. Drake, M. Joy, A. Goldenberg, and D. Kreindler, "Computer- and robot-assisted resection of thalamic astrocytomas in children," *Neurosurgery*, vol. 29, no. 1, pp. 27–33, Jul. 1991.

[94] L. R. Kavoussi, R. G. Moore, J. B. Adams, and A. W. Partin, "Comparison of Robotic Versus Human Laparoscopic Camera Control," *The Journal of Urology*, vol. 154, no. 6, pp. 2134–2136, Dec. 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022534701667156

[95] A. K. Sachdeva, "Surgical Education to Improve the Quality of Patient Care: the Role of Practice-Based Learning and Improvement," *J Gastrointest*

Surg, vol. 11, no. 11, pp. 1379–1383, Nov. 2007. [Online]. Available: http://link.springer.com.proxy1.library.jhu.edu/article/10.1007/s11605-007-0261-5

[96] H. Abboudi, M. S. Khan, O. Aboumarzouk, K. A. Guru, B. Challacombe, P. Dasgupta, and K. Ahmed, "Current status of validation for robotic surgery simulators a systematic review," *BJU Int*, vol. 111, no. 2, pp. 194–205, Feb. 2013. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1464-410X.2012.11270.x/abstract

[97] A. Moglia, V. Ferrari, L. Morelli, M. Ferrari, F. Mosca, and A. Cuschieri, "A Systematic Review of Virtual Reality Simulators for Robot-assisted Surgery," *European Urology*, vol. 69, no. 6, pp. 1065–1080, Jun. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S030228381500929X

[98] M. A. Liss, C. Abdelshehid, S. Quach, A. Lusch, J. Graversen, J. Landman, and E. M. McDougall, "Validation, Correlation, and Comparison of the da Vinci Trainer and the da Vinci Surgical Skills Simulator Using the Mimic Software for Urologic Robotic Surgical Education," *Journal of Endourology*, vol. 26, no. 12, pp. 1629–1634, Jul. 2012. [Online]. Available: http://online.liebertpub.com.proxy1.library.jhu.edu/doi/abs/10.1089/end.2012.0328

[99] S. P. DiMaio and C. Hasser, "The da Vinci research interface," *The MIDAS Journal - Systems and Architectures for Computer Assisted*

Interventions (MICCAI 2008 Workshop), Jul. 2008. [Online]. Available: http://www.midasjournal.org/browse/publication/622

[100] A. Deguet, R. Kumar, R. H. Taylor, and P. Kazhanzides, "The cisst libraries for computer assisted intervention systems," 2008. [Online]. Available: http://hdl.handle.net/10380/1465

[101] B. Vagvolgyi, S. P. Dimaio, A. Deguet, P. Kazhanzides, R. Kumar, C. Hasser, and R. Taylor, "The Surgical Assistant Workstation," *Systems and Architectures for Computer Assisted Interventions (MICCAI 2008 Workshop)*, 2008. [Online]. Available: http://hdl.handle.net/10380/1466

[102] H. C. Lin, "Structure in surgical motion," Ph.D., The Johns Hopkins University, United States – Maryland, 2010. [Online]. Available: http://search.proquest.com/dissertations/docview/365906486/abstract/142197187AD4114BBD/1?accountid=11752

[103] B. Varadarajan, "Learning and inference algorithms for dynamical system models of dextrous motion," Ph.D., The Johns Hopkins University, United States – Maryland, 2011. [Online]. Available: http://search.proquest.com/dissertations/docview/921360801/abstract/141E747011F77B39196/1?accountid=11752

[104] S. S. Vedula, A. O. Malpani, L. Tao, G. Chen, Y. Gao, P. Poddar, N. Ahmidi, C. Paxton, R. Vidal, S. Khudanpur, G. D. Hager, and C. C. G. Chen, "Analysis of the structure of surgical activity for a suturing and knot-tying

task," *PLOS ONE*, vol. 11, no. 3, p. e0149174, Mar. 2016. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0149174

[105] S. S. Vedula, A. Malpani, N. Ahmidi, S. Khudanpur, G. Hager, and C. C. G. Chen, "Task-level vs. segment-level quantitative metrics for surgical skill assessment," *J Surg Educ*, vol. 73, no. 3, pp. 482–489, May 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1931720415002901

[106] V. Datta, S. Mackay, A. Darzi, and D. Gillies, "Motion Analysis in the Assessment of Surgical Skill," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 4, no. 6, pp. 515–523, Jan. 2001. [Online]. Available: http://dx.doi.org/10.1080/10255840108908024

[107] A. Dosis, A. Aggarwal, F. Bello, K. Moorthy, Y. Munz, D. Gillies, and A. Darzi, "Synchronized video and motion analysis for the assessment of procedures in the operating theater," *Arch Surg*, vol. 140, no. 3, pp. 293–299, Mar. 2005. [Online]. Available: http://dx.doi.org/10.1001/archsurg.140.3.293

[108] J. Rosen, J. Brown, L. Chang, M. Sinanan, and B. Hannaford, "Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 399–413, 2006.

[109] B. Varadarajan, C. Reiley, H. Lin, S. Khudanpur, and G. Hager, "Data-

derived models for segmentation with application to surgical assessment and training," in *Medical Image Computing and Computer-Assisted Intervention*, G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, and C. Taylor, Eds. Springer Berlin Heidelberg, Jan. 2009, pp. 426–434. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-04268-3_53

[110] C. E. Reiley and G. D. Hager, "Task versus Subtask Surgical Skill Evaluation of Robotic Minimally Invasive Surgery," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2009*, ser. Lecture Notes in Computer Science, G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, and C. Taylor, Eds. Springer Berlin Heidelberg, Jan. 2009, no. 5761, pp. 435–442. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-04268-3_54

[111] C. E. Reiley, H. C. Lin, D. D. Yuh, and G. D. Hager, "Review of methods for objective surgical skill evaluation," *Surg Endosc*, vol. 25, no. 2, pp. 356–366, Jul. 2010. [Online]. Available: http://link.springer.com/article/10.1007/s00464-010-1190-z

[112] R. Kumar, G. D. Hager, A. S. Jog, Y. Gao, M. Liu, S. P. DiMaio, B. Itkowitz, and M. Curet, "Method and system for analyzing a task trajectory," U.S. Patent US20 140 378 995 A1, Dec., 2014, u.S. Classification 606/130; International Classification A61B5/06, A61B19/00; Cooperative Classification

A61B2034/107, A61B34/30, A61B19/2203, A61B5/065. [Online]. Available: http://www.google.com/patents/US20140378995

[113] J. Surowiecki, *The Wisdom of Crowds*. Knopf Doubleday Publishing Group, Aug. 2005, google-Books-ID: hHUsHOHqVzEC.

[114] W. S. Lasecki, Y. C. Song, H. Kautz, and J. P. Bigham, "Real-time Crowd Labeling for Deployable Activity Recognition," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, ser. CSCW '13. New York, NY, USA: ACM, 2013, pp. 1203–1212. [Online]. Available: http://doi.acm.org/10.1145/2441776.2441912

[115] A. Malpani, S. S. Vedula, C. C. G. Chen, and G. D. Hager, "A study of crowdsourced segment-level surgical skill assessment using pairwise rankings," *Int J CARS*, vol. 10, no. 9, pp. 1435–1447, Jun. 2015. [Online]. Available: http://link.springer.com/article/10.1007/s11548-015-1238-6

[116] E. Estells-Arolas and F. Gonzlez-Ladrn-de Guevara, "Towards an integrated crowdsourcing definition," *Journal of Information Science*, vol. 38, no. 2, pp. 189–200, Apr. 2012. [Online]. Available: http://jis.sagepub.com/content/38/2/189

[117] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960. [Online]. Available: http://epm.sagepub.com/content/20/1/37

[118] J. L. Fleiss, B. Levin, and M. C. Paik, "The measurement of interrater agreement," in *Statistical Methods for Rates and Proportions*. John Wiley & Sons, Inc., 2003, pp. 598–626. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/0471445428.ch18/summary

[119] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol Bull*, vol. 76, no. 5, pp. 378–382, 1971.

[120] A. Malpani, C. Lea, C. C. G. Chen, and G. D. Hager, "System events: readily accessible features for surgical phase detection," *Int J CARS*, vol. 11, no. 6, pp. 1201–1209, May 2016. [Online]. Available: http://link.springer.com/article/10.1007/s11548-016-1409-0

[121] J. D. Birkmeyer, J. F. Finks, A. O'Reilly, M. Oerline, A. M. Carlin, A. R. Nunn, J. Dimick, M. Banerjee, and N. J. Birkmeyer, "Surgical skill and complication rates after bariatric surgery," *New Engl J Med*, vol. 369, no. 15, pp. 1434–1442, Oct. 2013. [Online]. Available: http://www.nejm.org/doi/full/10.1056/NEJMsa1300625

[122] A. C. P. Gudon, M. Paalvast, F. C. Meeuwsen, D. M. J. Tax, A. P. van Dijke, L. S. G. L. Wauben, M. van der Elst, J. Dankelman, and J. J. van den Dobbelsteen, "It is Time to Prepare the Next patient Real-Time Prediction of Procedure Duration in Laparoscopic Cholecystectomies," *Journal*

of Medical Systems, vol. 40, no. 12, Dec. 2016. [Online]. Available: http://link.springer.com/10.1007/s10916-016-0631-1

[123] H. C. Lin, I. Shafran, D. Yuh, and G. D. Hager, "Towards automatic skill evaluation: detection and segmentation of robot-assisted surgical motions," *Comput. Aided Surg.*, vol. 11, no. 5, pp. 220–230, Sep. 2006.

[124] S.-K. Jun, M. Narayanan, P. Agarwal, A. Eddib, P. Singhal, S. Garimella, and V. Krovi, "Robotic Minimally Invasive Surgical skill assessment based on automated video-analysis motion studies," in *2012 4th IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2012, pp. 25–31.

[125] F. Despinoy, D. Bouget, G. Forestier, C. Penet, N. Zemiti, P. Poignet, and P. Jannin, "Unsupervised Trajectory Segmentation for Surgical Gesture Recognition in Robotic Training," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1280–1291, Jun. 2016.

[126] Y. Gao, S. S. Vedula, G. I. Lee, M. R. Lee, S. Khudanpur, and G. D. Hager, "Unsupervised surgical data alignment with application to automatic activity annotation," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 4158–4163.

[127] R. DiPietro, C. Lea, A. Malpani, N. Ahmidi, S. S. Vedula, G. I. Lee, M. R. Lee, and G. D. Hager, "Recognizing surgical activities

with recurrent neural networks," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2016*, ser. Lecture Notes in Computer Science, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Springer International Publishing, Oct. 2016, no. 9900, pp. 551–558, dOI: 10.1007/978-3-319-46720-7_64. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-46720-7_64

[128] B. B. Haro, L. Zappella, and R. Vidal, "Surgical gesture classification from video data," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2012*, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds. Springer Berlin Heidelberg, Jan. 2012, pp. 34–41. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-33415-3_5

[129] C. Lea, R. Vidal, and G. D. Hager, "Learning convolutional action primitives for fine-grained action recognition," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 1642–1649.

[130] C. Lea, A. Reiter, R. Vidal, and G. D. Hager, "Segmental Spatiotemporal CNNs for Fine-grained Action Segmentation," *arXiv:1602.02995 [cs]*, Feb. 2016, arXiv: 1602.02995. [Online]. Available: http://arxiv.org/abs/1602.02995

[131] C. Rupprecht, C. Lea, F. Tombari, N. Navab, and G. D. Hager, "Sensor Substitution for Video-based Action Recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016.

[132] L. Zappella, B. Bejar, G. Hager, and R. Vidal, "Surgical gesture classification from video and kinematic data," *Med Image Anal*, vol. 17, no. 7, pp. 732–745, Oct. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841513000522

[133] L. Tao, L. Zappella, G. D. Hager, and R. Vidal, "Surgical Gesture Segmentation and Recognition," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2013*, ser. Lecture Notes in Computer Science, K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, Eds. Springer Berlin Heidelberg, Jan. 2013, no. 8151, pp. 339–346. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-40760-4_43

[134] C. Lea, G. D. Hager, and R. Vidal, "An Improved Model for Segmentation and Recognition of Fine-Grained Activities with Application to Surgical Training Tasks," in *2015 IEEE Winter Conference on Applications of Computer Vision*, Jan. 2015, pp. 1123–1129.

[135] Y. Jin, Q. Dou, H. Chen, L. Yu, and P.-A. Heng, "EndoRCN: Recurrent Convolutional Networks for Recognition of Surgical Work ow in Cholecystectomy Procedure Video," *Modeling and Monitoring of Computer Assisted Interventions (M2CAI) Surgical Workflow Challenge*, 2016. [Online]. Available: http://camma.u-strasbg.fr/m2cai2016/reports/Jin-Workflow.pdf

[136] R. Cadne, T. Robert, N. Thome, and M. Cord, "M2cai Workflow Challenge:

Convolutional Neural Networks with Time Smoothing and Hidden Markov Model for Video Frames Classification," *arXiv:1610.05541 [cs]*, Oct. 2016, arXiv: 1610.05541. [Online]. Available: http://arxiv.org/abs/1610.05541

[137] A. P. Twinanda, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "Single- and Multi-Task Architectures for Surgical Workflow Challenge at M2cai 2016," *arXiv:1610.08844 [cs]*, Oct. 2016, arXiv: 1610.08844. [Online]. Available: http://arxiv.org/abs/1610.08844

[138] A. P. Twinanda, M. D. Mathelin, and N. Padoy, "Fisher Kernel Based Task Boundary Retrieval in Laparoscopic Database with Single Video Query," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2014*, ser. Lecture Notes in Computer Science, P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, Eds. Springer International Publishing, Sep. 2014, no. 8675, pp. 409–416. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-10443-0_52

[139] O. Dergachyova, D. Bouget, A. Huaulm, X. Morandi, and P. Jannin, "Automatic data-driven real-time segmentation and recognition of surgical workflow," *Int J CARS*, vol. 11, no. 6, pp. 1081–1089, Mar. 2016. [Online]. Available: http://link.springer.com/article/10.1007/s11548-016-1371-x

[140] T. Blum, H. Feussner, and N. Navab, "Modeling and segmentation of surgical workflow from laparoscopic video," *Med Image Comput Comput Assist Interv*,

vol. 13, no. Pt 3, pp. 400–407, 2010. [Online]. Available: http://campar.in.tum.de/pub/blum2010miccaiWorkflow/blum2010miccaiWorkflow.pdf

[141] S.-A. Ahmadi, T. Sielhorst, R. Stauder, M. Horn, H. Feussner, and N. Navab, "Recovery of Surgical Workflow Without Explicit Models," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2006*, ser. Lecture Notes in Computer Science, R. Larsen, M. Nielsen, and J. Sporring, Eds. Springer Berlin Heidelberg, Oct. 2006, no. 4190, pp. 420–428, dOI: 10.1007/11866565_52. [Online]. Available: http://link.springer.com/chapter/10.1007/11866565_52

[142] N. Padoy, T. Blum, I. Essa, H. Feussner, M. O. Berger, and N. Navab, "A boosted segmentation method for surgical workflow analysis," *Med Image Comput Comput Assist Interv*, vol. 10, no. Pt 1, pp. 102–109, 2007.

[143] N. Padoy, T. Blum, S.-A. Ahmadi, H. Feussner, M.-O. Berger, and N. Navab, "Statistical modeling and recognition of surgical workflow," *Medical Image Analysis*, vol. 16, no. 3, pp. 632–641, Apr. 2012. [Online]. Available: http://www.medicalimageanalysisjournal.com/article/S1361841510001131/abstract

[144] R. Stauder, A. Okur, L. Peter, A. Schneider, M. Kranzfelder, H. Feussner, and N. Navab, "Random Forests for Phase Detection in Surgical Workflow Analysis," in *Information Processing in Computer-Assisted Interventions*, ser. Lecture Notes in Computer Science, D. Stoyanov, D. L. Collins,

I. Sakuma, P. Abolmaesumi, and P. Jannin, Eds. Springer International Publishing, Jun. 2014, no. 8498, pp. 148–157. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-07521-1_16

[145] R. DiPietro, R. Stauder, E. Kayis, A. Schneider, M. Kranzfelder, H. Feussner, G. D. Hager, and N. Navab, "Automated Surgical-Phase Recognition Using Rapidly-Deployable Sensors," *Modeling and Monitoring of Computer Assisted Interventions (M2CAI)*, 2015. [Online]. Available: https://www.iccas.de/m2cai2015/index.html

[146] D. Katic, A.-L. Wekerle, F. Gartner, H. Kenngott, B. P. Muller-Stich, R. Dillmann, and S. Speidel, "Knowledge-Driven Formalization of Laparoscopic Surgeries for Rule-Based Intraoperative Context-Aware Assistance," in *Information Processing in Computer-Assisted Interventions*, ser. Lecture Notes in Computer Science, D. Stoyanov, D. L. Collins, I. Sakuma, P. Abolmaesumi, and P. Jannin, Eds. Springer International Publishing, Jun. 2014, no. 8498, pp. 158–167. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-07521-1_17

[147] D. Kati, C. Julliard, A.-L. Wekerle, H. Kenngott, B. P. Mller-Stich, R. Dillmann, S. Speidel, P. Jannin, and B. Gibaud, "LapOntoSPM: an ontology for laparoscopic surgeries and its application to surgical phase

recognition," *Int J CARS*, vol. 10, no. 9, pp. 1427–1434, Jun. 2015. [Online]. Available: http://link.springer.com/article/10.1007/s11548-015-1222-1

[148] D. Kati, J. Schuck, A.-L. Wekerle, H. Kenngott, B. P. Mller-Stich, R. Dillmann, and S. Speidel, "Bridging the gap between formal and experience-based knowledge for context-aware laparoscopy," *Int J CARS*, vol. 11, no. 6, pp. 881–888, Mar. 2016. [Online]. Available: http://link.springer.com/article/10.1007/s11548-016-1379-2

[149] G. Forestier, L. Riffaud, and P. Jannin, "Automatic phase prediction from low-level surgical activities," *Int J CARS*, vol. 10, no. 6, pp. 833–841, Apr. 2015. [Online]. Available: http://link.springer.com/article/10.1007/s11548-015-1195-0

[150] T. Neumuth, P. Jannin, G. Strauss, J. Meixensberger, and O. Burgert, "Validation of knowledge acquisition for surgical process models," *J Am Med Inform Assoc*, vol. 16, no. 1, pp. 72–80, Feb. 2009.

[151] C. Meiner, J. Meixensberger, A. Pretschner, and T. Neumuth, "Sensor-based surgical activity recognition in unconstrained environments," *Minimally Invasive Therapy & Allied Technologies*, vol. 23, no. 4, pp. 198–205, Aug. 2014. [Online]. Available: http://www.tandfonline.com/doi/abs/10.3109/13645706.2013.878363

[152] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn*,

vol. 20, no. 3, pp. 273–297, Sep. 1995. [Online]. Available: http://link.springer.com/article/10.1007/BF00994018

[153] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: http://link.springer.com/article/10.1023/A%3A1010933404324

[154] S. Sarawagi and W. W. Cohen, "Semi-Markov Conditional Random Fields for Information Extraction," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT Press, 2005, pp. 1185–1192. [Online]. Available: http://papers.nips.cc/paper/2648-semi-markov-conditional-random-fields-for-information-extraction.pdf

[155] G. Navarro, "A Guided Tour to Approximate String Matching," *ACM Comput. Surv.*, vol. 33, no. 1, pp. 31–88, Mar. 2001. [Online]. Available: http://doi.acm.org/10.1145/375360.375365

[156] R. Kwitt, S. Hegenbart, N. Rasiwasia, A. Vcsei, and A. Uhl, "Do We Need Annotation Experts? A Case Study in Celiac Disease Classification," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2014*, ser. Lecture Notes in Computer Science, P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, Eds. Springer International Publishing, Jan. 2014, no. 8674, pp. 454–461. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-10470-6_57

[157] A. Malpani, S. S. Vedula, C. C. G. Chen, and G. D. Hager, "Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task," in *Information Processing in Computer-Assisted Interventions*, D. Stoyanov, D. L. Collins, I. Sakuma, P. Abolmaesumi, and P. Jannin, Eds. Springer International Publishing, Jan. 2014, pp. 138–147. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-07521-1_15

[158] J. L. Cameron, "William Stewart Halsted. Our surgical heritage." *Ann Surg*, vol. 225, no. 5, pp. 445–458, May 1997. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1190776/

[159] R. Reznick, G. Regehr, H. MacRae, J. Martin, and W. McCulloch, "Testing technical skill via an innovative bench station examination," *The American Journal of Surgery*, vol. 173, no. 3, pp. 226–230, Mar. 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002961097895979

[160] B. Law, M. S. Atkins, A. E. Kirkpatrick, and A. J. Lomax, "Eye Gaze Patterns Differentiate Novice and Experts in a Virtual Laparoscopic Surgery Training Environment," in *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*, ser. ETRA '04. New York, NY, USA: ACM, 2004, pp. 41–48. [Online]. Available: http://doi.acm.org/10.1145/968363.968370

[161] N. Ahmidi, G. D. Hager, L. Ishii, G. Fichtinger, G. L. Gallia, and M. Ishii, "Surgical Task and Skill Classification from Eye Tracking and

Tool Motion in Minimally Invasive Surgery," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2010*, ser. Lecture Notes in Computer Science, T. Jiang, N. Navab, J. P. W. Pluim, and M. A. Viergever, Eds. Springer Berlin Heidelberg, Sep. 2010, no. 6363, pp. 295–302, dOI: 10.1007/978-3-642-15711-0_37. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-15711-0_37

[162] N. Ahmidi, M. Ishii, G. Fichtinger, G. L. Gallia, and G. D. Hager, "An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool-motion data," *International Forum of Allergy & Rhinology*, vol. 2, no. 6, pp. 507–515, Nov. 2012. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/alr.21053/abstract

[163] Y. R. Naikavde, "Robotic Surgical Skill Assessment Based on Pattern Classification Tools," M.S., State University of New York at Buffalo, United States – New York, 2012. [Online]. Available: http://search.proquest.com/dissertations/docview/1112501227/abstract/141B2B454F71D10CE31/1?accountid=11752

[164] S.-k. Jun, M. S. Narayanan, P. Singhal, S. Garimella, and V. Krovi, "Evaluation of robotic minimally invasive surgical skills using motion studies," *J Robotic Surg*, vol. 7, no. 3, pp. 241–249, Sep. 2013. [Online]. Available: http://link.springer.com/article/10.1007/s11701-013-0419-y

[165] N. Ahmidi, P. Poddar, J. D. Jones, S. S. Vedula, L. Ishii, G. D. Hager,

and M. Ishii, "Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty," *Int J CARS*, pp. 1–11, Apr. 2015. [Online]. Available: http://link.springer.com/article/10.1007/s11548-015-1194-1

[166] A. Khan, S. Mellor, E. Berlin, R. Thompson, R. McNaney, P. Olivier, and T. Pltz, "Beyond Activity Recognition: Skill Assessment from Accelerometer Data," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '15. New York, NY, USA: ACM, 2015, pp. 1155–1166. [Online]. Available: http://doi.acm.org/10.1145/2750858.2807534

[167] H. Rafii-Tari, C. Payne, J. Liu, C. Riga, C. Bicknell, and G.-Z. Yang, "Towards automated surgical skill evaluation of endovascular catheterization tasks based on force and motion signatures," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 1789–1794.

[168] L. Richstone, M. J. Schwartz, C. Seideman, J. Cadeddu, S. Marshall, and L. R. Kavoussi, "Eye Metrics as an Objective Assessment of Surgical Skill:," *Annals of Surgery*, vol. 252, no. 1, pp. 177–182, Jul. 2010. [Online]. Available: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00000658-201007000-00028

[169] T. Tien, P. H. Pucher, M. H. Sodergren, K. Sriskandarajah, G.-Z. Yang,

and A. Darzi, "Eye tracking for skills assessment and training: a systematic review," *Journal of Surgical Research*, vol. 191, no. 1, pp. 169–178, Sep. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022480414004326

[170] E. D. Gomez, R. Aggarwal, W. McMahan, K. Bark, and K. J. Kuchenbecker, "Objective assessment of robotic surgical skill using instrument contact vibrations," *Surg Endosc*, pp. 1–13, Jul. 2015. [Online]. Available: http://link.springer.com/article/10.1007/s00464-015-4346-z

[171] R. Kumar, A. Jog, A. Malpani, B. Vagvolgyi, D. Yuh, H. Nguyen, G. Hager, and C. Chen, "Assessing system operation skills in robotic surgery trainees," *Int J Med Rob Comput Assist Surg*, vol. 8, no. 1, pp. 118–124, 2012. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/rcs.449/abstract

[172] I. Nisky, Y. Che, Z. F. Quek, M. Weber, M. Hsieh, and A. Okamura, "Teleoperated versus open needle driving: Kinematic analysis of experienced surgeons and novice users," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 5371–5377.

[173] M. K. Powers, A. Boonjindasup, M. Pinsky, P. Dorsey, M. Maddox, L.-m. Su, M. Gettman, C. P. Sundaram, E. P. Castle, J. Y. Lee, and B. R. Lee, "Crowdsourcing assessment of surgeon dissection of renal artery and vein during robotic partial nephrectomy: A novel approach for quantitative

assessment of surgical performance," *Journal of Endourology*, Nov. 2015. [Online]. Available: http://online.liebertpub.com.proxy1.library.jhu.edu/doi/abs/10.1089/end.2015.0665

[174] L. L. Thurstone, "A law of comparative judgment," *Psychological Review*, vol. 34, no. 4, pp. 273–286, 1927.

[175] R. Kumar, P. Rajan, S. Bejakovic, S. Seshamani, G. Mullin, T. Dassopoulos, and G. Hager, "Learning disease severity for capsule endoscopy images," 2009, pp. 1314–1317.

[176] Y. Freund, R. Iyer, R. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *J Mach Learn Res*, vol. 4, pp. 933–969, Dec. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=945365.964285

[177] A. E. Elo, *The rating of chessplayers, past and present.* Arco Pub., 1978.

[178] R. Herbrich, T. Minka, and T. Graepel, "TrueSkill(TM): A Bayesian Skill Rating System," in *Advances in Neural Information Processing Systems 20.* MIT Press, Jan. 2007, pp. 569–576. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=67956

[179] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," ser. WWW '01. New York, NY, USA: ACM, 2001, pp. 613–622. [Online]. Available: http://doi.acm.org/10.1145/371920.372165

[180] A.-L. D. D'Angelo, D. N. Rutherford, R. D. Ray, S. Laufer, C. Kwan, E. R. Cohen, A. Mason, and C. M. Pugh, "Idle time: an underdeveloped performance metric for assessing surgical skill," *The American Journal of Surgery*, vol. 209, no. 4, pp. 645–651, Apr. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002961015000045

[181] C. C. G. Chen, E. Tanner, A. Malpani, S. S. Vedula, A. N. Fader, S. A. Scheib, I. C. Green, and G. D. Hager, "Warm-up before robotic hysterectomy does not improve trainee operative performance: a randomized trial," *J Minim Invasive Gynecol*, vol. 22, no. 6, Supplement, p. S34, Nov. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1553465015007013

[182] R. S. Wigton, K. D. Patil, and V. L. Hoellerich, "The effect of feedback in learning clinical diagnosis," *J Med Educ*, vol. 61, no. 10, pp. 816–822, Oct. 1986.

[183] E. Salas, K. A. Wilson, C. S. Burke, and H. A. Priest, "Using simulation-based training to improve patient safety: what does it take?" *Jt Comm J Qual Patient Saf*, vol. 31, no. 7, pp. 363–371, Jul. 2005.

[184] D. A. Rogers, G. Regehr, T. R. Howdieshell, K. A. Yeh, and E. Palm, "The impact of external feedback on computer-assisted learning for surgical technical skill training," *The American Journal of Surgery*,

vol. 179, no. 4, pp. 341–343, Apr. 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S000296100000341X

[185] T. Mahmood and A. Darzi, "The learning curve for a colonoscopy simulator in the absence of any feedback: No feedback, no learning," *Surg Endosc*, vol. 18, no. 8, pp. 1224–1230, Jun. 2004. [Online]. Available: http://link.springer.com/article/10.1007/s00464-003-9143-4

[186] E. Boyle, D. A. O'Keeffe, P. A. Naughton, A. D. K. Hill, C. O. McDonnell, and D. Moneley, "The importance of expert feedback during endovascular simulator training," *Journal of Vascular Surgery*, vol. 54, no. 1, pp. 240–248.e1, Jul. 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0741521411002242

[187] J. Oestergaard, F. Bjerrum, M. Maagaard, P. Winkel, C. R. Larsen, C. Ringsted, C. Gluud, T. Grantcharov, B. Ottesen, and J. L. Soerensen, "Instructor feedback versus no instructor feedback on performance in a laparoscopic virtual reality simulator: a randomized educational trial," *BMC Med Educ*, vol. 12, p. 7, Feb. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3311079/

[188] D. Stefanidis, J. R. Korndorffer Jr., B. T. Heniford, and D. J. Scott, "Limited feedback and video tutorials optimize learning and resource utilization during laparoscopic simulator training," *Surgery*, vol. 142, no. 2, pp. 202–206, Aug.

2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0039606007002644

[189] F. Bjerrum, M. Maagaard, J. Led Sorensen, C. Rifbjerg Larsen, C. Ringsted, P. Winkel, B. Ottesen, and J. Strandbygaard, "Effect of Instructor Feedback on Skills Retention After Laparoscopic Simulator Training: Follow-Up of a Randomized Trial," *Journal of Surgical Education*, vol. 72, no. 1, pp. 53–60, Jan. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1931720414001913

[190] C. W. Snyder, M. J. Vandromme, S. L. Tyra, and M. T. Hawn, "Proficiency-Based Laparoscopic and Endoscopic Training With Virtual Reality Simulators: A Comparison of Proctored and Independent Approaches," *Journal of Surgical Education*, vol. 66, no. 4, pp. 201–207, Jul. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1931720409001123

[191] L. B. Rosenberg, "Virtual fixtures: Perceptual tools for telerobotic manipulation," in *, 1993 IEEE Virtual Reality Annual International Symposium, 1993*, Sep. 1993, pp. 76–82.

[192] S. S. Vedula, K. Olds, M. Balicki, G. Gallia, G. D. Hager, R. H. Taylor, and M. Ishii, "Robot-assisted active learning (RAAL) of surgical technical skill," London, UK, Jun. 2016. [Online]. Available: http://hamlyn.doc.ic.ac.uk/hsmr/sites/default/files//HSMR-proceedings_FINAL.pdf

[193] Institute of Medicine (US) Committee on Quality of Health Care in America, *To Err is Human: Building a Safer Health System*, L. T. Kohn, J. M. Corrigan, and M. S. Donaldson, Eds. Washington (DC): National Academies Press (US), 2000. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK225182/

[194] A. Gardner and M. Rich, "Error management training and simulation education," *Clin Teach*, vol. 11, no. 7, pp. 537–540, Dec. 2014. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/tct.12217/abstract

[195] M. A. Fischer, K. M. Mazor, J. Baril, E. Alper, D. DeMarco, and M. Pugnaire, "Learning from Mistakes," *J Gen Intern Med*, vol. 21, no. 5, pp. 419–423, May 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1484785/

[196] A. King, M. G. Holder, and R. A. Ahmed, "Errors as allies: error management training in health professions education," *BMJ Qual Saf*, pp. bmjqs–2012–000 945, Jan. 2013. [Online]. Available: http://qualitysafety.bmj.com/content/early/2013/01/03/bmjqs-2012-000945

[197] S. Arora, R. Aggarwal, P. Sirimanna, A. Moran, T. Grantcharov, R. Kneebone, N. Sevdalis, and A. Darzi, "Mental Practice Enhances Surgical Technical Skills: A Randomized Controlled Study," *Annals of Surgery*, vol. 253, no. 2, pp. 265–270, Feb. 2011. [Online]. Available: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00000658-201102000-00009

[198] S. J. Cole, H. Mackenzie, J. Ha, G. B. Hanna, and D. Miskovic, "Randomized controlled trial on the effect of coaching in simulated laparoscopic training," *Surg Endosc*, vol. 28, no. 3, pp. 979–986, Nov. 2013. [Online]. Available: http://link.springer.com/article/10.1007/s00464-013-3265-0

[199] K. A. Ericsson, "Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains," *Acad Med*, vol. 79, no. 10 Suppl, pp. S70–81, Oct. 2004.

[200] ——, "The Influence of Experience and Deliberate Practice on the Development of Superior Expert Performance," in *The Cambridge handbook of expertise and expert performance*, K. A. Ericsson, N. Charness, P. J. Feltovich, and R. R. Hoffman, Eds. New York, NY, US: Cambridge University Press, 2006, pp. 683–703.

[201] K. A. Ericsson, R. T. Krampe, and C. Tesch-Rmer, "The role of deliberate practice in the acquisition of expert performance," *Psychological Review*, vol. 100, no. 3, pp. 363–406, Jul. 1993. [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=1993-40718-001&site=ehost-live&scope=site

[202] D. W. Birch, A. H. Asiri, and C. J. de Gara, "The impact of a formal mentoring program for minimally invasive surgery on surgeon practice and patient outcomes," *The American Journal of Surgery*,

vol. 193, no. 5, pp. 589–592, May 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002961007000554

[203] J. M. Brit, M. J. E. Mourits, M. J. A. Kenkhuis, A. G. J. van der Zee, G. H. de Bock, and H. J. G. Arts, "Implementing an Advanced Laparoscopic Procedure by Monitoring with a Visiting Surgeon," *Journal of Minimally Invasive Gynecology*, vol. 17, no. 6, pp. 771–778, Nov. 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1553465010002736

[204] Y.-Y. Hu, S. E. Peyre, A. F. Arriaga, R. T. Osteen, K. A. Corso, T. G. Weiser, R. S. Swanson, S. W. Ashley, C. P. Raut, M. J. Zinner, A. A. Gawande, and C. C. Greenberg, "Postgame Analysis: Using Video-Based Coaching for Continuous Professional Development," *Journal of the American College of Surgeons*, vol. 214, no. 1, pp. 115–124, Jan. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1072751511011604

[205] S. Palmer, "PRACTICE: A model suitable for coaching, counselling, psychotherapy and stress management." *The Coaching Psychologist*, vol. 3, no. 2, pp. 72–77, Aug. 2007. [Online]. Available: https://www.researchgate.net/publication/294888712_Palmer_S_2007_PRACTICE_A_model_suitable_for_coaching_counselling_psychotherapy_and_stress_management_The_Coaching_Psychologist_3_2_72-77

[206] M. D. Karam, G. W. Thomas, D. M. Koehler, B. O. Westerlind, P. M. Lafferty,

G. T. Ohrt, J. L. Marsh, A. E. V. Heest, and D. D. Anderson, "Surgical Coaching from Head-Mounted Video in the Training of Fluoroscopically Guided Articular Fracture Surgery," *J Bone Joint Surg Am*, vol. 97, no. 12, pp. 1031–1039, Jun. 2015. [Online]. Available: http://jbjs.org/content/97/12/1031

[207] B. Renshaw and G. Alexander, *Super Coaching: The Missing Ingredient for High Performance*. Random House, 2005, google-Books-ID: VLmuAluvwi4C.

[208] V. N. Palter, K. A. Beyfuss, A. R. Jokhio, A. Ryzynski, and S. Ashamalla, "Peer coaching to teach faculty surgeons an advanced laparoscopic skill: A randomized controlled trial," *Surgery*, Jun. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0039606016301246

[209] M. L. Soucisse, K. Boulva, L. Sideris, P. Drolet, M. Morin, and P. Dub, "Video Coaching as an Efficient Teaching Method for Surgical ResidentsA Randomized Controlled Trial," *Journal of Surgical Education*, Oct. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1931720416301568

[210] Greenberg CC, Dombrowski J, and Dimick JB, "Video-based surgical coaching: An emerging approach to performance improvement," *JAMA Surg*, vol. 151, no. 3, pp. 282–283, Mar. 2016. [Online]. Available: http://dx.doi.org/10.1001/jamasurg.2015.4442

[211] D. Stefanidis, B. Anderson-Montoya, R. V. Higgins, M. E. Pimentel, P. Rowland, M. O. Scarborough, and D. Higgins, "Developing a coaching

mechanism for practicing surgeons," *Surgery*, vol. 160, no. 3, pp. 536–545, Sep. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0039606016301088

[212] D. M. Rooney, E. S. Hungness, D. A. DaRosa, and C. M. Pugh, "Can skills coaches be used to assess resident performance in the skills laboratory?" *Surgery*, vol. 151, no. 6, pp. 796–802, Jun. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0039606012001195

[213] "Open Source Haptics - H3d.org." [Online]. Available: http://h3dapi.org/

[214] M. Nathan, J. M. Karamichalis, H. Liu, S. Emani, C. Baird, F. Pigula, S. Colan, R. R. Thiagarajan, E. A. Bacha, and P. del Nido, "Surgical technical performance scores are predictors of late mortality and unplanned reinterventions in infants after cardiac surgery," *The Journal of Thoracic and Cardiovascular Surgery*, vol. 144, no. 5, pp. 1095–1101.e7, Nov. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022522312009282

[215] S. E. Regenbogen, C. C. Greenberg, D. M. Studdert, S. R. Lipsitz, M. J. Zinner, and A. A. Gawande, "Patterns of Technical Error Among Surgical Malpractice Claims: An Analysis of Strategies to Prevent Injury to Surgical Patients," *Ann Surg*, vol. 246, no. 5, pp. 705–711, Nov. 2007.

[216] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin,

L. Tao, L. Zappella, B. Bejar, D. Yuh, C. C. G. Chen, R. Vidal, S. Khudanpur, and G. D. Hager, "The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling," in *Modeling and Monitoring of Computer Assisted Interventions (M2CAI) MICCAI Workshop, 2014*, Sep. 2014.

[217] "Guidelines for Academic Requesters - WeAreDynamo Wiki." [Online]. Available: http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters

[218] P. G. Ipeirotis, "Analyzing the Amazon Mechanical Turk Marketplace," *XRDS*, vol. 17, no. 2, pp. 16–21, Dec. 2010. [Online]. Available: http://doi.acm.org/10.1145/1869086.1869094

[219] L. C. Irani and M. S. Silberman, "Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13. New York, NY, USA: ACM, 2013, pp. 611–620. [Online]. Available: http://doi.acm.org/10.1145/2470654.2470742

Vita

Anand Malpani was born in 1989 in Mumbai, India. He received his B.Tech. in Electrical Engineering at Indian Institute of Technology (IIT) Bombay in 2010. He undertook a summer research project in medical imaging at the Insitut de Recherche en Communications et Cybernétechnique de Nantes under the guidance of Vincent Ricordel in 2009. He joined the Ph.D. program in Computer Science at Johns Hopkins University in 2010. His dissertation under the guidance of Gregory D. Hager, focused on surgical education and simulation-based training with the goal of automated surgical coaching. He was awarded the Intuitive Surgical Student Fellowship in 2013, and the Link Foundation's Modeling, Training and Simulation Fellowship in 2015. He was a summer research intern in the Simulation team at Intuitive Surgical Inc. (Sunnyvale, CA) in 2015. He will be an Assistant Research Scientist at the Malone Center for Engineering in Healthcare at JHU starting March 2017.